1995

Differential Item Functioning in Performance Assessments: A Comparison of Three Procedures.

Hae-seong Park
Louisiana State University and Agricultural & Mechanical College

DIFFERENTIAL ITEM FUNCTIONING IN PERFORMANCE ASSESSMENTS: A COMPARISON OF THREE PROCEDURES

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Administrative and Foundational Services

by
Hae-Seong Park
B.A. Chongshin College, 1980
M.Div. Chongshin Seminary, 1983
M.C.E. Reformed Theological Seminary, 1989
Th.M. Reformed Theological Seminary, 1990
December 1995

ACKNOWLEDGEMENTS

It would be impossible to acknowledge explicitly all those individuals who have differentially functioned in favor of me in this endeavor. However, I would like to try to express my gratitude in words to some of those individuals.

Before I begin to thank all of the many wonderful people, I have to first praise the Lord's almighty grace. Like the meaning of my name (i.e., HAE-SEONG in Korean means "Success in the Grace of Lord"), my Lord has touched my life and has led my way. In fact, I know that He has worked and will work in all things of my life for good.

I wish to thank my parents who have been so patient, so understanding, and so supportive with prayer. To my wife I give a special thanks, the most wonderful woman in this world, who has always believed in me and has sacrificed so much so that I may pursue what has been so important to me.

I would like to express my sincere gratitude to Dr. Eugene Kennedy, who served as my committee chairperson throughout this dissertation. Without his inspiration and guidance, this dissertation would not have been possible. It was under his stimulating tutelage that I was formally introduced to the field of psychometrics in general and the Differential Item Functioning in particular.

I am especially grateful to Dr. Abbas Tashakkori whose contributions to my advanced education in general and statistics in particular are tremendous and immeasurable. He always responded with great patience and care no matter what

kind of questions I asked or how shallow the questions were. I know that he was the most DIF person in my graduate studies.

It is noteworthy that Dr. Charles Teddlie was the individual who first offered me a substantial research assistanship which enable me to pursue my doctoral studies at Louisiana State University and who recommended me to the Louisiana Department of Education for experience and practice of the professional career. I know that he is an invisible hand to whom I really owe a very special debt of gratitude.

I also appreciate Dr. Kim MacGregor who served not only as a member on my committee but also as the individual who opened my eyes in the area of educational technology.

I wish to express my gratitude to Dr. Scott M. Norton, my friend and mentor, whose interrogations and suggestions greatly improved the quality of the manuscript.

Finally, I wish to extend a very special thanks to all members in Lafayette Korean Church, old brothers and sisters in the Starkville Korean Church, and all of the staff in Bureau of Pupil Accountability at the Louisiana Department of Education. Their love, support, and belief in me has kept me going.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

As performance assessments grow in popularity, it becomes increasingly important to investigate the effect of such assessments on various population subgroups. The purpose of this study was to investigate the relative empirical power of three popular statistical procedures (an extension of the generalized Mantel-Haenszel procedure, Logistic Discriminant Function Analysis, and a combined *t*-test procedure) in identifying polytomously scored items that function differentially for two subgroups of examinees.

In the Monte Carlo study computer simulations were conducted to study the behavior of these procedures for identifying items exhibiting varying degrees of differential-item functioning (DIF). Each statistic was converted to a probability value to examine the number of times that the method rejected an item at the .05 levels.

The results, based on simulated twenty-four conditions, each replicated 50 times, indicate a preference for the logistic discriminant function analysis (LDFA) procedure for DIF identification in polytomously scored items. The effects of the number of DIF items on the matching variable seem significant for identifying DIF in performance assessment. The effect was stronger for detecting uniform DIF than for identifying nonuniform DIF.

Based on the findings of the study, the following conclusions were drawn:

(1) For DIF analysis in performance assessments, the LDFA can be recommended as the preferred method to test constructors or practitioners. (2) Through using the LDFA for identifying DIF in performance assessments, the appropriateness in test usage for different subgroups will be enlarged. (3) The effects of the number of DIF items on the matching variable seem significant for identifying DIF in performance assessment. Thus, in order to decrease the effects of the proportion of DIF items on the matching variable, it is recommended to emphasize the judgmental analysis to evaluate biased items in a test before entering DIF analysis.

Finally, the statistics should be interpreted with caution. Although DIF analysis is essential for the appropriateness of test use that is related to subgroups influenced by testing, DIF analysis is only one component of the extensive research for the validity and fairness of performance assessment.

CHAPTER ONE: INTRODUCTION

Overview

General Discussion of Historical Concerns about Testing

From elementary and secondary school, to the college admission process, to professional employment, the capacity to perform well on tests influences the lives of people. Today testing is the prime vehicle in the quest for fair selection, based solely on ability. Unfortunately, the history of testing is filled with examples that cause concern about unfairness and bias in test construction and test use.

Testing is not a modern creation. Numerous types of tests have been used throughout time and across cultures. The ancient Spartans tested the prowess of their youth in the martial arts (Harman, 1980). The rabbis of the ancient academies of Babylon assessed the intellectual abilities of their students. In many cultures, ceremonies of initiation involved tests of various kinds. Tests of knowledge and ability were frequently employed in apprenticeships.

The use of mental tests seems to be as old as Western civilization itself. In the Bible, there is a short verbal test, a kind of performance assessment, depicted when Jephthah uses the term "Shibboleth" as a test word by which to distinguish the fleeing Ephraimites from his own Gileadites. As Wainer (1990) mentions, although this test could have resulted in death, there was no validity study.

Some formal testing occurred in China around 2200 B.C. in which written examination papers were rewritten to eliminate one possible source of bias in the grading of the exams (Popham, 1990). This testing program was modified through the years and was advocated for use in France and England by the beginning of the nineteenth century (Wainer, 1990).

Universities lagged far behind in their efforts to install examination systems. The first exam which was given at the University of Bologna in 1219 was exclusively an oral exam (Popham, 1990). The tradition of oral exams spread quickly and written exams, also used by the middle of the nineteenth century, were widely applied in the United States and Western Europe. By the beginning of the twentieth century, serious research efforts began on the use of various testing procedures.

The wave of activity in testing at the beginning of the twentieth century reached a broader range of disciplines than just psychology. The most significant contribution was from statistics, when Spearman provided the rudiments of psychometrics- reliability coefficients (Wainer, 1990). A major change in test practice from individualized to mass administrated occurred at this time.

An influential pioneer in the testing movement was Joseph Mayer Rice who studied methods of augmenting the efficiency of schooling in the late 1880s. By administering his tests to large samples of school children and establishing the

average scores to be expected at different grade levels, Rice's work contributed heavily to early thinking about the use of standardized tests in education (Popham, 1990). Rice's efforts greatly influenced E. L. Thorndike who not only refined some of Rice's approaches to measurement but also evolved a host of important technical advances (Popham, 1990).

The first major experiment in group intelligence testing, which originated from Binet's intelligence scales, took place during World War I when psychologists were involved in the implementation of a program for the psychological examination of recruits. These researchers developed the Army Alpha, which was administered to approximately 2,000,000 recruits, and the Army Beta, which was for non-English speaking recruits (Wainer, 1990). Together, the two forms represented the first large-scale use of intelligence testing.

During the years after World War I, psychologists applied their skills in testing to the civilian world (i.e., industry and schools). By 1926, the success of testing programs within the military had influenced the College Board (Wigdor & Garner, 1982). The College Board introduced the *Scholastic Aptitude Test* which became a major tool in admission decisions and scholarship competition, particularly at the most prestigious colleges.

As the technology for creating valid tests matured, their use broadened to include industrial placement and advancement tests. During this time, test companies such as Educational Testing Service (ETS), American College Testing (ACT), and the Psychological Cooperation (Psych Corp), were founded. Testing in industry, spurred on by the availability of inexpensive tests from private test producers and from the U.S. Employment Service, increased just as dramatically. By the 1960s, almost all American businesses, of at least moderate size, were using some kind of employment test (Wigdor & Garner, 1982).

Emerging Concern for Fairness in Testing

In recent years much attention has been directed to the issue of fairness in educational and psychological testing. The issue of fairness in testing was highlighted during the Civil Rights Movement of the 1960s (Pulliam, 1991) and further emphasized during the Women's Rights movement that followed.

The following tables, which were taken from a recent report of a statewide criterion-referenced testing program, show different performance patterns between all minority and majority groups and male and female students. All students represented in this table failed to pass one or more subjects of the criterion-referenced test, and as a result they could not obtain the high school diploma. According to the statistics of the state, the public school student

population consists of 51 percent white, 49 percent African-American, 50 percent male, and 50 percent female students.

Table 1.1 Number and Percentage of Seniors of Each Ethnic Group Who Did Not Attain One or More Subjects in Graduation Examination

Ethnicity	Number	Percentage (%)
African-American	**1,153**	**77.59**
White	**268**	**18.03**
Hispanic	27	1.82
Asian-American	28	1.88
Native-American	3	0.20
Missing (invalid)	7	0.47
Total	1,486	100.00

Table 1.2 Number and Percentage of Seniors of Each Gender Group Who Did Not Attain One or More Subjects in Graduation Examination

Ethnicity	Number	Percentage (%)
Male	**577**	**38.83**
Female	**904**	**60.83**
Missing (invalid)	5	0.34
Total	1,486	100.00

Concerns about fairness and equal rights for certain groups raised the need to examine equity in testing, particularly as it relates to access to educational and professional opportunities (McAllister, 1993). Of course, the issue of fairness involves both test use and test construction. Conceivably, a test might be unbiased, meaning that examinees from different groups were not unfairly penalized by test content or administration, but the use of the test could be unfair

to one or more groups. For example, if members of one group performed better on a measure used to select individuals for a training program than members of another group, and this fact was reflected in test scores, the test is certainly not biased. However, if test scores have no association with performance in the training program, use of the test as a selection mechanism is unfair to persons in the low-scoring group.

In the late 1960s and 1970s, concern about fairness focused almost exclusively on Intelligence Quotient (IQ) tests, which were used in education to make tracking and special education placement decisions (Jensen, 1980). IQ tests were used extensively for employee selection and placement in many industries until the 1971 *Griggs v. Duke Power Company* case (Camilli and Shepard, 1994). This verdict stated that tests which ended in disproportionate hiring by ethnicity must not be used unless employers could show a direct relationship between the tests and job performance.

During this period, argument surrounding the issue of test fairness was expanded by reaction to the publication of Jensen's (1969) article on the heritability of intelligence in the *Harvard Educational Review*. Jensen argued that IQ is influenced much more by genetic factors than by environmental effects. This conclusion has significant political meaning, because it implies that observed differences in group performance are the result of genetic factors rather

than past discrimination. One of the responses to Jensen was the position statement of the Association of Black Psychologists calling for a suspension on all testing of African-Americans until more equitable tests became available (Camilli & Shepard, 1994).

The politically-charged issue of test fairness in the 1960s and 1970s led to many studies. One of the first studies was by Cleary (1968). She defined unbiased and biased tests using the regression model for predicting performance from test scores of examinees in minority and majority groups. She defined it as follows:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance (p. 115).

Following this definition, Thorndike (1971) also presented definitions of test bias. Thorndike suggested that a test is unbiased if the minority and majority groups are as far apart on predictor scores as they are on criterion scores. Darlington (1971) discussed four definitions of test bias. He maintained that no single test can meet all the specifications likely to be made for a *culturally* fair or

unbiased test. Darlington's important point about these definitions is that if a test is unbiased by one definition, it is almost certain to be biased by the other definitions. For example, Thorndike's definition, which Peterson and Novick (1976) refer to as the Constant Ratio Model, conflicts with Cleary's definition. These and other results have led reseachers to conclude that the issue of bias in selection is largely a question of sociopolitical values. For example, no matter how the regression lines of minority and majority groups compare, a selection procedure can be plotted to fit any particular sociopolitical purpose (Crocker & Algina, 1986). The full spectrum of test validity and fairness issues cannot be captured by any one technical-bias model.

In the mid-1970s, item-bias procedures received considerable attention because it was convenient to analyze the items within a test when external criteria were not available. Methodologically, the design of internal-item bias analyses was to distinguish between true group differences and bias in the measurement system. Group differences on test items could not be defined automatically as an indication of bias because score differences might be valid reflections of group differences in knowledge. Therefore, the concept of "relative difficulty" (Camilli & Shepard, 1994) was devised. When no external criterion was available, a variety of internal bias procedures were developed using the total test score as the criterion for determining real group differences. If the statistical result from these internal bias procedures shows relative difficulty for a specific group, a

second step is needed to conclude that the item is biased. This step is to determine whether the relative difficulty is irrelevant to the test construct or not. If an item is relatively more difficult for one group, and the source of this difficulty is irrelevant to the test construct, the item is biased.

The term *differential item functioning* (DIF) is now widely used in the literature. To maintain the distinction between relative difficulty and bias, psychometricians refer to unexplained relative difficulty as differential-item functioning (DIF) (Holland & Thayer, 1988). DIF statistics are used to identify all items that function differently for different groups. After qualitative review or logical analysis as to why the items seem to be relatively more difficult for a particular group, a subset of DIF items would be identified as "biased." Thus, an item demonstrates DIF or potential bias if examinees of equal ability, but from different subgroups, do not have equal probability of correctly responding to that item. Figure 1.1 shows the conceptual framework of the process for identifying biased test items.

As shown in Figure 1.1, in DIF analysis extensive statistical analyses are accomplished to identify the relative performance of major subgroups of examinees on individual test items. If any items are detected as DIF items, they will be submitted for judgmental analysis at the next step. The remaining items

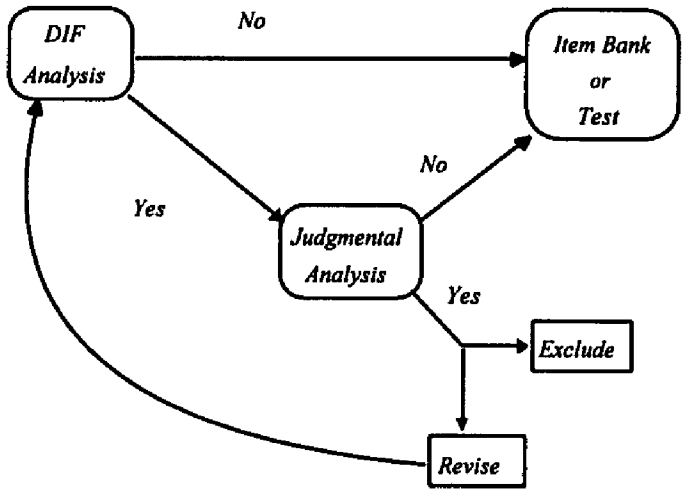

Figure 1.1 The Conceptual Framework of the Item Bias Method

will be stored in an item bank. In this second step, details of DIF items will be reviewed by subject-matter experts and members of the major subgroups in society that will be represented in the examinee population. When items are identified as biased items through this second analysis, they will be excluded or revised. The items which are not identified as biased items by the committee will be stored in the item bank. However, the items which are revised by test developers still need to be examined again to determine if they are free from bias using additional DIF analyses.

Bias can best be understood within the context of test validity. The concept of test validity has evolved over time. The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1985) state that validity "refers

to the appropriateness, meaningfulness, and usefulness of the specific inference made from test scores (p. 9)." Messick (1989) noted that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (p.13)." Based on these statements, validity can be best defined as the extent to which certain inferences can be made accurately from test scores.

One of the issues involving threats to validity is multidimensionality. Fundamentally, DIF is a symptom of multidimensionality among groups. Some test items may simply function differently for members from one group or another, or they may measure different constructs for examinees from one group or another. The existence of such differential item functioning implies that something other than the property intended to be measured influences performance on the item. Thus, the item reflects more than one dimension of individual difference variation. The validity of inferences drawn from such a test is clouded because resulting scores may be an indication of a variety of attributes other than those the test is purposed to measure.

In the process of detecting DIF, several methods [e.g., transformed item difficulty (TID) index, adjustments to the TID index, and ANOVA] were developed based on classical test theory. However, because of flaws in these

indices (See Chapter 2), most are no longer recommended (Camilli & Shepard, 1994). Recent approaches can be divided into those based on item response theory and contingency-table based procedures (Hambleton, Clauser, Mazor, & Jones, 1993).

Performance Assessment as an Alternative Assessment

In addition to concerns with test bias, the past decade has witnessed increased scrutiny of conventional objective measures. Until recently, testing practices were primarily based on objective measures such as multiple-choice, true-false and matching items. However, alternatives to objective measures have rapidly become a prime topic of discussion at the national, state and local levels (Wiggins, 1993). The proponents of alternative assessment argue that objective tests have a negative impact on teaching and learning, penalize creativity and are racially or culturally biased (Wiggins, 1989) These critics go on to argue in favor of assessments that are authentic in the sense that they (a) are consistent with current knowledge about teaching and learning; (b) promote creativity and respect diversity; and (c) provide direct, not indirect, measures of desired skills (Wiggins, 1993).

One type of alternative or authentic assessment is performance assessment. Stiggins (1991) described performance assessment as a mark of the beginning of a new era of assessment. According to Stiggins (1991),

performance assessment is "the observation and rating of student behavior and products in contexts where students actually demonstrate proficiency (p. 273)." Performance assessment is increasingly being incorporated into tests in the United States through the addition of a practical component to traditional multiple-choice tests. These components are called performance items and require examinees to perform a practical task. Examples of nationwide testing and assessment programs that include performance items are the College Board Advanced Placement tests, the National Assessment of Educational Progress (NAEP) writing, reading, science, and mathematics assessments, the Praxis Series (successor to the NTE teacher assessment), the ACT College Outcome Measures Program and Workkeys assessments (Zwick, Donoghue, & Grima, 1993). Many States in the United States are also administering or planning pure performance assessments in areas such as writing and reading.

According to a recent report state assessment programs (NCREL, 1994), forty-two of forty-seven states surveyed have created non-traditional test items. Of the five states who have not created non-traditional test items, two states plan to develop such items. In particular, the report shows that written composition or reading tests are administered in a statewide criterion-referenced test by thirteen states.

In the excitement that surrounds the new assessment methodologies, attention must still focus on the continuing requirements for accurate educational and psychological measurement. All educational and psychological assessments of students' abilities should satisfy professional measurement standards (as exemplified in the 1985 *Standards for Educational and Psychological Testing*), regardless of the testing and measurement method used. Therefore, performance assessments must also demonstrate the following: (1) sufficient reliability to support the selection or classification of individuals, (2) validity to support inferences concerning the achievements, aptitudes, and performance capabilities of those assessed, (3) fairness or an unbiased way to reflect the abilities of those assessed without regard to gender, ethnic group membership, or socio-economic status, and (4) support for the classification of examinees into decision-relevant categories (Hambleton, 1994).

Unfortunately, although the belief has been expressed that performance assessments are substantially more fair than multiple-choice measures, some forms of performance assessment may be more likely than conventional tests to produce construct-irrelevant factors (Zwick, Donoghue, & Grima, 1993). For example, in a written composition test, when item responses are scored by raters who know the identity of each respondent or who can guess the respondent's gender or ethnicity, rater bias can occur. If respondents tend to receive higher

scores from raters of their own ethnicity, then respondents who are scored by same-ethnicity raters will have an unfair advantage. Clearly, adding performance sections to an existing test might lead to larger mean differences among ethnic groups (Dunbar, Koretz, & Hoover, 1991). This larger mean difference may occur either due to group differences in the construct or due to construct-irrelevant factors.

Performance Assessment and Differential Item Functioning

In contrast to conventional objective items which are usually dichotomously scored (yes/no, correct/incorrect), the scoring of many performance assessments involves a range of proficiency. For example, a student's proficiency at some task might be scored as 1 (incompetent), 2 (competent), or 3 (superior). Items of this type are called polytomously-scored items and as performance assessments grow in popularity, it will become increasingly important to investigate the validity and fairness of these item formats. Specifically, either new procedures must be developed or current procedures for dichotomously scored items (e.g., the Mantel-Haenszel common-odds ratio, logistic regression) must be generalized to the polytomously scored items.

An extension of the Mantel-Haenszel procedure that may be useful in assessing DIF for polytomously-scored items was proposed by Zwick, Donohue,

and Grima (1993). However, there is no single extension of the logistic regression procedure to accommodate polytomously-scored items. Several approaches can be followed, such as the model of the response probabilities pairwise using adjacent categories, and the model of several logistic regression analyses with recoded polytomous responses and cumulative logit models for fitting sequences of cumulative probabilities. However, these approaches require assumptions that may not be warranted in a DIF analysis, such as the equal-slopes regression lines assumption. Therefore, the logistic-discriminant function analysis was proposed by Miller and Spray (1993). They argued that this method is capable of handling any type of item response with more flexibility than existing methods. Also, it is very effective in detecting the existence of nonuniform DIF. It should be noted that nonuniform DIF exists when the discrepancies in the probabilities of a right answer for the two groups are not consistent across all ability levels, while uniform DIF exists when there is a relative advantage for one group over the entire ability range for an item.

Welch and Hoover (1993) also recently proposed the combined *t*-test method for use in detecting DIF in polytomously scored items. Desirable features of this method appear to be the ease of calculating and interpreting the statistics associated with DIF.

For the purposes of this study, there are two main streams of performance assessment: a test in which performance items are included with conventional objective items, and a test that consists of all performance items-- a pure performance test. Recently-proposed DIF methods for polytomously-scored items have been studied almost exclusively for the first case. For example, Zwick, Donoghue and Grima's (1993) study had 20 dichotomous items and 4 polytomous items, Miller and Spray's (1993) study consisted of 21 dichotomous items and 6 polytomous items, and Welch and Hoover's (1993) study included all dichotomous items as a matching variable. These studies have relevance because many current applications of performance assessments incorporate both conventional and performance items. However, some believe that the future direction for performance assessment is likely to be pure performance assessment, exclusive of dichotomous items (Calfee & Perfumo, 1993). Therefore, there is a need to study DIF in performance assessment using a matching criterion which consists of only polytomous items.

Lastly, as discussed in the definition of *Differential Item Functioning*, DIF methods rely on internal criteria. Therefore, it should be assumed that the total test score is unbiased. However, if a biased item exists in a test, the total test score cannot be considered free from bias. One method to correct this problem is to exclude the studied item in computing the total test score or matching criterion.

Then, if there is more than one biased item in a test, the problem will be more complicated. However, in a practical testing situation, it cannot be guaranteed that only a few items are biased. In fact, internal DIF methods are incapable of detecting constant bias. Thus, there is a need to study the effect of the number of DIF items on the matching variable to identify DIF in performance assessment.

According to a Monte Carlo study by Donoghue and Others (1993) for dichotomously-scored items, a relatively small influence of the number of DIF items on the matching variable was found . Unfortunately, the effect of the number or proportion of DIF items on the matching variable for performance assessment has not been explored.

Statement of the Problem

Current trends indicate that performance assessments will increase in prominence in national, state, and local testing programs in the future. The enthusiasm for these procedures exhibited by measurement specialists and practitioners alike belies the fact that, due to the recentness of their popularity, the research is limited. In particular, there is the possibility that these assessments will exacerbate problems of test bias. Furthermore, statistical procedures for investigating bias or differential-item functioning for polytomously-scored items, the format most common for performance

assessments, are recent in development. As a consequence, their properties and performance are not well understood.

The Purpose of the Study

The purpose of this study is to investigate, through Monte-Carlo procedures, the relative efficacy or power of three popular statistical procedures (an extension of the generalized Mantel-Haenszel procedure, Logistic Discriminant Function Analysis, and a combined *t*-test procedure) for investigating DIF for polytomously-scored test items. The performance of these procedures will be studied under a variety of experimental conditions.

Significance/Importance of the Study

This study is significant because it investigates, in a single study, three recently-proposed methods that appear to be promising in the assessment of DIF in polytomously-scored items. A significant contribution of this study is to provide information about which method is more powerful than the others in a variety of different conditions. For test constructors or practitioners, the results of this study will provide guidance in selecting and implementing proper DIF procedures for performance-assessment data. Also, this study will provide data about the effect of the proportion of biased polytomously-scored items in a test on DIF detection rates.

As discussed above, the basic purpose of a DIF study is to detect items which are disadvantageous to subgroups of a population. Although the DIF process is only one component of the extensive research that is needed on the validity and fairness of performance assessments, this process is an essential element. Therefore, this study ultimately contributes towards a greater appropriateness in test usage for groups affected by testing.

Research Questions

The following questions were developed to examine the relative efficiency or power of three statistical procedures for identifying DIF in performance assessment. These questions were divided into two categories. The first group of questions were generated to compare the relative statistical powers of three methods based on the integer transformed theta (ITT) score as the matching variable. The ITT, which will be discussed in detail in Chapter 3, was created to make a criterion that is free from DIF. The second group of questions were developed to determine the effect of the number of DIF items on the matching variable for detecting DIF in polytomously scored items. The design of the study to answer the second group of questions was based on the total test scores as the matching variable.

Category I

Research question 1.

Which statistical method is the most powerful when there are differences in ability between Reference and Focal groups?

Research question 2.

Which statistical method is the most powerful when the size of the sample is relatively small?

Research question 3.

Which statistical method is the most powerful when there are unequal sample sizes between Reference and Focal groups?

Research question 4.

Which statistical method is the most powerful when nonuniform DIF exists?

Research question 5.

Which statistical method demonstrates consistent control of Type I error under the null hypothesis?

Research question 6.

Which statistical method is the most powerful for detecting DIF across all conditions?

Category II

Research question 7.

Is there any effect of the proportion of DIF in a test on detecting DIF?

Research question 8.

If any effect of the proportion of DIF items in a test exists on detecting DIF, which statistical method is the most efficient for this condition?

After determining the most powerful statistical method for identifying performance items that are disadvantageous to subgroups of a population, a preferable method will be proposed. As a consequence of using this method, the fairness of test or test validity for groups influenced by testing will be greatly improved. However, it should be noted that DIF analysis is only one step in a complex process for determining the validity and fairness of performance assessments.

Limitations of the Study

This study only deals with three DIF assessing procedures--an extension of the Mantel-Haenszel procedure (Zwick, Donoghue, & Grima, 1993), Logistic Discriminant Function Analysis (Miller & Spray, 1993), and a combined *t*-test procedure (Welch & Hoover, 1993). Although item response theory (IRT), which will be discussed in detail in Chapter 2, provides powerful and flexible techniques for investigating DIF, the Mantel-Haenszel procedure, the logistic

regression procedure (contingency table approaches), and the combined *t*-test procedure often do as well as IRT methods (Camilli & Shepard, 1994; Welch & Hoover, 1993). Because proper implementation and interpretation of IRT methods require considerable sophistication, the three methods may be preferable in application where users lack the technical knowledge or computer resources to implement IRT methods (Camilli & Shepard, 1994).

For the purposes of this study, a three-step partial credit model (Masters & Wright, 1984) will be used to describe the performance items with possible score values of 1 to 4. However, there are many types of polytomous items in performance assessment. For instance, possible score values of 1 to 3 (no control, moderate control, and high control, or disagree, undecided, and agree) and possible score values of 1 to 5 (strongly disagree, disagree, undecided, agree, and strongly disagree) are also currently used in performance assessment. The results discussed here may not generalize to these situations.

In this study, only simulation data will be used. As noted above, the use of simulation data instead of actual data allows for greater variation of conditions. However, although some conditions are useful for examining the relative statistical power of the methods studied, they may have limited relevance in practical situations.

Definition of Terms

Test Bias

In order to define *test bias*, it is necessary to distinguish between *test bias* and *test fairness*. Test bias is a concern related to bias within a test, while test fairness is a concern about the way a test is used. Test bias has been defined as invalidity or systematic error in how a test measures any definable, relevant subgroup of test takers (Camilli and Shepard, 1994; Linn, 1989).

Item Bias

In order to define *item bias*, it is necessary to distinguish *item bias* from *test bias*. Westers (1993) argued that if a test is biased this does not mean that all the items of the test are biased. Generally, when a test is biased one or a few of the items are biased. However, if a test is not biased, the items in the test can still be biased, because it might be the case that bias in one item is compensated for by the bias of another item.

As discussed earlier, a biased item is a subset of DIF items. Therefore, a biased item can be defined as an item which is relatively more difficult for a certain group, and the source of this difficulty is irrelevant to the test construct.

Differential Item Functioning

In order to maintain the distinction between relative difficulty and bias, psychometricians refer to the unexplained relative difficulty as DIF (Holland &

Thayer, 1988). DIF statistics are used to identify all items that function differently for different groups. An item demonstrates DIF or potential bias if examinees of equal ability, but from different subgroups, do not have equal probability of correctly responding to that item. Thus, DIF is defined as differences in item functioning on groups that have been matched with respect to the ability that the item purportedly measures (Dorans & Holland, 1993). In this study if an item exhibits a statistically-significant unequal probability of a correct answer at the .05 level, it is defined as a DIF item.

Item Difficulty

Item difficulty is defined as an index which represents the degree of difficulty of an item. The difficulty of an item is defined operationally as the proportion of examinees (*p*-values) in a given population who answered the item correctly.

Item Discrimination

Item discrimination has been defined as "the relationship between the difficulty of an item and the ability of the examinees" (Osterlind, 1989; p. 282). Thus, item discrimination is usually considered as a relationship between a given item against the total score on the test itself because the total test score is used as an operational definition of the examinee's ability. Therefore, item

discrimination is operationally defined as a correlation between examinees' scores of an item and their total test scores.

Test Validity

Messick (1989) noted that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (p.13)." Based on these statements, validity can be best defined as the extent to which certain inferences can be made accurately from test scores.

Performance Assessment

Stiggins (1991) has defined performance assessment as "the observation and rating of student behavior and products in contexts where students actually demonstrate proficiency (p. 273)." Airasian (1991) defined performance assessment in which the test administrator observes and makes a judgment about an examinee's skill in carrying out an activity or producing a product. He distinguished performance assessment from essay tests and oral questions. In this study, however, various types of assessment which can be polytomously scored will be treated as performance assessment.

Polytomously Scored Items

The recent emphasis on performance assessment has raised the need for psychometric procedures that can address item responses other than those scored as correct/incorrect. Miller and Spray (1993) stated that item responses which are scored on a nominal or ordinal scale and which consist of more than two categories are termed polytomous item responses. The responses can be rated and scored through a scoring rubric, with partial credit given for phases or steps toward the solution. In this study, an item in which responses consist of four possible categories is treated as a polytomously scored item.

Statistical Power

The power of a statistical test is defined as "the probability that the test will correctly reject a false null hypothesis (Gravetter & Wallnau, 1992)." In this study, the power of each procedure was measured by estimating the probability that DIF items were detected as DIF items at the .05 level.

Uniform and Nonuniform DIF

Uniform DIF exists if there is no interaction between ability level and group membership. It means that the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability. Therefore, when there is a relative advantage for one group over the entire ability range for an item, the DIF is defined as uniform DIF.

Nonuniform DIF exists if there is interaction between ability level and group membership. It means that the discrepancies in the probabilities of a right answer for the two groups are not consistent across all ability levels.

Reference and Focal Groups

Usually the performance of two groups of examinees is compared in a DIF analysis. The group of primary interest is defined as the *focal group* and the other group which is used as a basis of comparison is called the *reference group*. In this study, the *focal* and *reference groups* are randomly generated from the sample of simulation data.

Matching Variables

The comparability of the reference and focal groups is achieved by matching them on the basis of a measure of test performance. Typically, this measure is the total score on the test to be evaluated for DIF and is defined as a matching variable (Schmitt, Holland, & Dorans, 1993). In this study, two criterion variables, the ITT score which is estimated from a simulated examinee's latent trait ability (θ) and the total test score, are employed as the matching variables.

Assumptions

This study was based on several assumptions which are discussed below. The first assumption was made when a simulation data set was utilized in this

study. Although the simulation may not accurately reflect reality in actual testing conditions, it is ideal for the current purpose. It should be noted that the results of this simulation study might be altered when less-than-ideal conditions exist.

The second assumption was brought about when the Integer Transformed Theta (ITT) scores were used as the matching variable. The ITT created a criterion that was free from DIF, however, this condition would be very difficult to obtain using real data.

The third assumption concerned the design of nonuniform DIF. A nonuniform DIF condition, in which positive and negative DIF cancel each other entirely, was designed for examining the relative statistical power of the three statistics. However, this ideal nonuniform DIF condition might be rare in practice.

The fourth assumption is associated with the effect of the number of DIF items in the matching variable. When the condition of the average magnitude of DIF items in a given test is the same as in the second test, the overall effect size of the DIF items on the total test score was directly proportional to the number of DIF items in a test. For example, it was assumed that there is no difference between the effect of two DIF items of 0.15 and 0.25 magnitudes and the effect of two items of 0.20 and 0.20 magnitudes.

Summary

This chapter provided a brief overview of the present study associated with the issues of test bias and performance assessment. Based on these two issues, this study was designed to investigate the relative efficiency of three statistics (i.e., the logistic discriminant function analysis, the Mantel-Haenszel procedure, and the combined *t*-test procedure) for identifying DIF in polytomously scored items. In order to examine the relative statistical power in a variety of different conditions, a simulation design was implemented. In the next chapter, a literature review of DIF procedures will be presented.

CHAPTER TWO. LITERATURE REVIEW

Overview

When items in educational or psychological tests show differential item functioning (DIF), the tests may be unfair for certain subgroups. It is important to identify these items so that they can be improved or removed from the tests. Although many DIF detection methods have been proposed, all available procedures are not discussed in this review. Rather, those chosen are the most prominent in the current literature (Angoff, 1993; Camilli & Shepard, 1994; Holland & Wainer, 1993). Four classes of methods are discussed: early bias indices based on classical test theory (transformed item difficulty (TID) method and analysis of variance (ANOVA)), contingency table approaches (Mantel-Haenszel procedure and logistic regression), item response theory (IRT) approaches, and standardization approach.

As noted in Chapter 1, the early bias indices based on classical test theory (TID and ANOVA) have technical flaws (Camilli & Shepard, 1994). However, a review of these methods serves several important purposes. Although they had technical defects, the logic of these indices provides a useful conceptual framework for the more complex and technically preferred methods. Also, a historical understanding of why previous conceptions were rejected explains why certain methods work better. For this reason, the first method, Transformed Item

Difficulty method, will be discussed more extensively than the other methods.

In the last section of this chapter three DIF assessing methods, an extension of Mantel-Haenszel procedure, logistic discriminant function analysis, and a combined t-test procedure for performance tasks will be discussed.

Classical Test Theory Approaches

The Transformed Item Difficulty Index (the Delta Plot Method)

In 1972, Angoff offered a method which is called variously the *delta-plot* or *transformed item-difficulty* (TID) method. This method became highly popular since it was computationally easy and intuitively reasonable.

In classical test theory, the difficulty of a test item is measured by the proportion of examinees getting the item correct. For the *g*th item, delta is defined as $\Delta_g = 4\, z_g + 13$. Here, z_g is the normalized z-score corresponding to the "(1 - p_g)th percentile", and p_g is the proportion-correct measure of item difficulty. To apply the delta plot procedure, the Δ_g's are calculated for each item and each subgroup. As illustrated in Figure 2.1, paired transformed item difficulties are arrayed in a bivariate plot reflecting the correspondence between the difficulty of the items in Group 1 and Group 2. If all items had exactly the same relative difficulty in both groups, the item data points would form a straight 45-degree line from the lower-left to the upper-right hand corner of the plot. The items falling at some distance from the plot of points, as measured by the distance

of the item's bivariate point from the principal axis of the plot, may be regarded as contributing to the item × group interaction (Angoff & Ford, 1973).

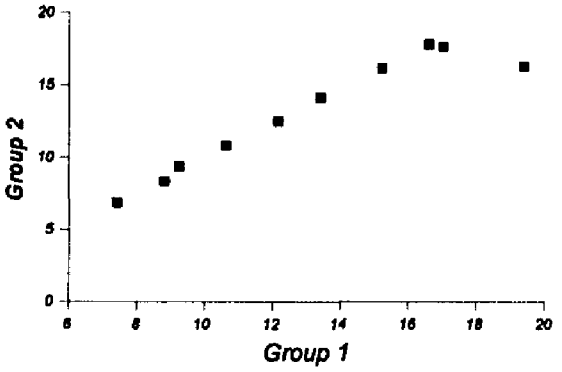

Figure 2.1 Transformed Item Difficulty

Table 2.1 presents hypothetical data describing the responses of samples from two subgroups to 10 items. A statistical method for detecting atypical items based on the TID index involves fitting a straight line to the scatterplot and calculating d_g, the distance of the gth item from the line. The line fitted to the scatterplot is the major axis of the plot. The absolute value of the distance of gth item from the major axis of the ellipse is:

$$d_g = \frac{k\Delta_{g1} - \Delta_{g2} + m}{\sqrt{k_2 + 2}} \qquad (2,1)$$

when k is the slope, and m is the intercept of the line. Items with large d_g's deviate sufficiently from the line to be considered biased.

Table 2.1. Comparison of Item Difficulty Indices for Two Subgroups

Item	p_{g1}	z_{g1}	Δ_{g1}	p_{g2}	z_{g2}	Δ_{g2}	d_g
1	.35	-.59	10.64	.39	-.54	10.85	-.27
2	.46	-.21	12.17	.49	-.12	12.50	-.47
3	.78	.90	16.61	.81	1.20	17.80	-1.52
4	.25	-.94	9.25	.30	-.91	9.36	-.10
5	.98	1.60	19.39	.72	.83	16.31	2.62
6	.55	.10	13.42	.59	.29	14.16	-.93
7	.68	.56	15.22	.71	.79	16.15	-1.19
8	.12	-1.39	7.44	.15	-1.53	6.87	.66
9	.22	-1.04	8.83	.24	-1.16	8.36	.49
10	.81	1.01	17.03	.80	1.16	17.64	-.96

This data was created by Hae-Seong Park through SAS program for demonstrating TID.

Table 2.1 reports the values of d_g for ten items. Inspection of the table shows that d_3 = -1.52 and d_5 = 2.62. These values are substantially larger than the values of d_g for the remaining items, which suggests that these items are biased. However, there does not seem to be a rule for deciding when d_g is large enough to indicate bias. Therefore, if several of the d_g's are not obviously larger than the others, interpretation is more difficult.

Despite its easy logic and practical simplicity, the TID approach may be flawed as an indicator of differential item functioning. As pointed out by Cole and Moss (1989), Angoff (1982), and Camilli and Shepard (1994), unless all the items have the same discriminating power the method may yield misleading results, especially when the groups obtain mean scores at widely different ability levels. Whenever two groups are not equal on the trait being measured, highly

discriminating items will appear to be biased because they do a better job of making the distinction between low-scoring and high-scoring groups.

Table 2.2 shows a simple example to demonstrate this problem.

Table 2.2. Comparison of Different Item Discrimination

	Group 1 p-value	Group 2 p-value	Difference in ps
Item A	.35	.75	.40
Item B	.40	.50	.10

In this case, item A would appear to be biased against Group 1 because it shows a much bigger difference between the two groups than does Item B. When there are many items in a test like Item B, the TID method will identify Item A as biased against Group 1 (See Figure 2.2).

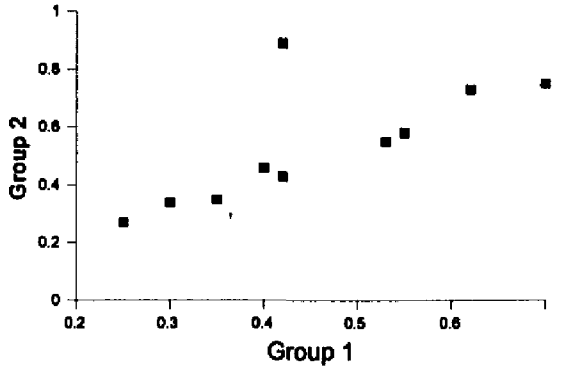

Figure 2.2 Discrimination Index and Item Bias (Detected a High Discriminating Item)

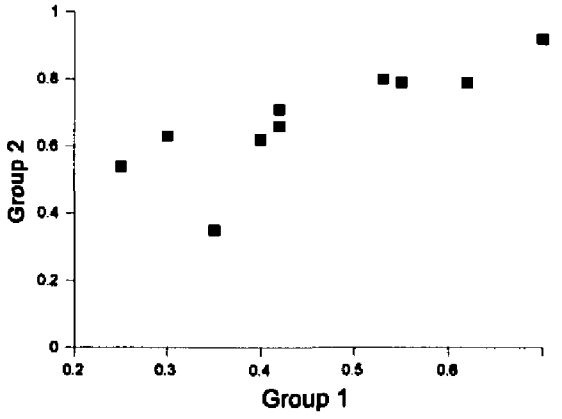

Figure 2.3 Discrimination Index and Item Bias (Detected a Low Discriminating Item)

In contrast, when there are many items in the test like Item A, Item B will be identified as biased against Group 2 because it shows a much smaller discrepancy than the typical difference between the two groups (See Figure 2.3). Therefore, item *p*-values cannot be trusted as valid indicators of differential item functioning. Whenever two groups are not equal on the trait being measured, large item discrepancies can readily occur as a function of item discrimination even when all items are measuring in precisely the same way for all groups.

Angoff (1982) suggested adjustments to the TID method to correct the limitations mentioned above, specifically, to match groups on a relevant external criterion prior to conducting a TID procedure. However, in practical applications, external criterion data are usually not available. Also, he recommended that the external criterion should be examined and declared free of

bias to be utilized as a matching variable. He suggested using an adjustment based on item-test correlations (point biserials). Angoff divided each item's transformed *p*-value by its point biserial correlation before constructing the plot of delta values. However, the point biserial correlations are unreliable, so the adjustments might be unreliable, too. Moreover, Shepard, Camilli, and Williams (1985) found that the adjustment TID procedure actually reduced the consistency of the TID index with the item response theory approach and chi-square indices. Thus, the TID method cannot be recommended as a means to detect differential item functioning (DIF).

Analysis of Variance(ANOVA) Method

The analysis of variance technique for detecting item bias is another method based on the *p*-value. ANOVA is a statistical method for analyzing variance into a number of additive components which equal the total score variance. A two-factorial, repeated measures ANOVA is conducted with examinee group membership as one factor, and items as the within-subjects factor. The total score variance can be written as:

$$\sigma^2_t = \sigma^2_g + \sigma^2_i + \sigma^2_{gi} + \sigma^2_e$$

where σ^2_t is the total variance, σ^2_g is the variance due to group, σ^2_i is the variance due to items, σ^2_{gi} is the variance of the interaction of items and groups, and σ^2_e is the error variance.

Cleary and Hilton (1968) stated that an item × group interaction exists (H_1: $\sigma^2_{gi} \neq 0$) when all the items in the test do not maintain the same relative difficulties in both the major and minor groups. They assumed that average group differences are reflected in the main effect for groups, whereas differential difficulty is expected to be reflected in the item × group interaction effect (See Figure 2.4).

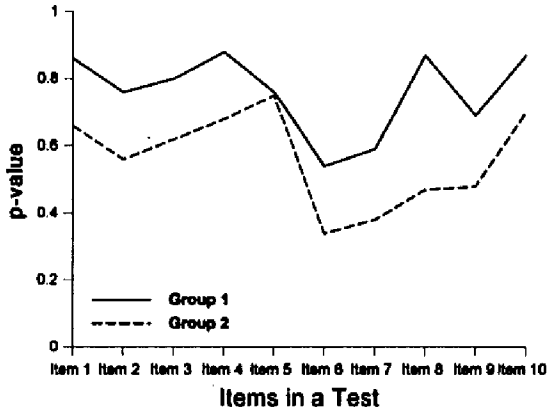

Figure 2.4 Interaction Effect in ANOVA Method (Group by Item)

The ANOVA method was the dominant method for analyzing internal test bias through the early 1980s. One of the most criticized shortcomings in ANOVA is that item bias by that technique is attributed only to a significant interaction. However, the interaction may be the effects of several variance sources. For instance, item score variance could be due to differences in item difficulty, the unequal ability of groups, or the student differences within groups. ANOVA

determines only whether variance components of interaction are significant and does not partition out the source of the variance (See Osterlind, 1983). Hunter (1975) found that item × group interaction effects would occur in completely unbiased tests because of the unequal ability of groups as noted in the TID method. Jensen (1980) concluded that ANOVA showed no evidence of test bias. He argued that "Analysis of variance reveals that the variance due to race × items interaction is very small (and often nonsignificant) as compared with the overall mean difference between the races (p. 586)". In a simulation study, Camilli and Shepard (1987) found that when the true mean group effect was greater than the true bias effect, more of the bias effect would be blended by the main effect than would appear in the interaction effect.

In summary, ANOVA method produces Type I error or Type II error-- that is, it will miss real occurrences of bias (Camilli & Shepard, 1987) as well as create artifactual bias (Hunter, 1975). Thus the ANOVA method can no longer be recommended as a bias-detection procedure.

Contingency Table Approaches

Early Chi-Square Techniques

In this section, two early chi-square techniques to detect DIF are briefly discussed. The first, Scheuneman's chi-square statistic ($\chi^2_{correct}$), focuses on

correct responses to a given item. The second, Camilli's chi-square statistic (χ^2_{full}), includes both correct and incorrect responses in the analysis.

According to Scheuneman (1979), "an unbiased item is defined as an item for which the probability of a correct response is the same for all persons of a given ability, regardless of their ethnic group membership" (p. 145). Baker (1981) criticized Scheuneman's procedure because it is based only on correct responses to an item. However, Camilli (Shepard, Camilli, & Averill, 1981) adopted a χ^2_{full} procedure, which incorporates both correct and incorrect responses in the calculation of the chi-square index. Both methods, $\chi^2_{correct}$ and χ^2_{full}, involve only two steps: (1) the observed score scale is divided into several intervals, and (2) the chi-square value is then calculated and tested for statistical significance.

A critical issue in the use of these chi-square methods is the choice of ability intervals. The first step in performing any chi-square procedure is to divide the total score scale into ability intervals or score levels. This step must be done with care and full recognition that expected frequencies in the chi-square statistic, and hence significance testing can be artificially altered by manipulation of the score intervals selected.

In general, most of the difficulty in setting intervals comes from the necessity of having an adequate number of observations from each group in the

upper and lower ability intervals. About 10 to 20 observed correct responses per cell is a recommended minimum in any case (See Scheuneman, 1979; Ironson, 1982). Typically, this results in about 3 to 5 ability categories. Ironson (1982) pointed out that several factors, including the overlap of the groups on the total score scale, the relative sizes of the groups, the difficulty of the item, and whether $\chi^2_{correct}$ or χ^2_{full} are to be performed, complicate selection of ability intervals. It is important that incorrect responses also be considered in selecting ability intervals. When no attention is given to the frequency of incorrect responses, the high extreme categories can become compressed often with just one or two score points comprising the entire ability interval.

Crocker and Algina (1986) pointed out that "one problem with the chi-square techniques is that evidence of item bias may be an artifact of measurement errors" (p. 386). The chi-square methods compare the subpopulations based on the proportion responding correctly to each item, but the comparison is made within each score interval. If the comparison were not made within each score interval, the result would be a comparison of the proportion of correct responses related to item difficulty. However, a subpopulation difference in item difficulty is not necessarily an indication of item bias. It may reflect true differences between the subgroups on the construct that the test is intended to measure.

Another problem with chi-square procedures is that they test against the H_0 but do not provide a measure of the amount of DIF (the magnitude of DIF) exhibited by the item. It is well known that tests always reject the null hypothesis when they have a large enough sample size. Therefore, it is more informative to have a measure of the size of the departure of the data from the null hypothesis. The Mantel-Haenszel procedure which will be discussed next provides such a measure.

Mantel-Haenszel Procedure

The Mantel-Haenszel (MH) statistical procedure was developed by Mantel and Haenszel (1959) and was applied to DIF study by Holland and Thayer (1988). On a *K*-item test of right-wrong items, the basic data used by the MH procedure are in the form of 2 (Groups/Reference or Focal) × 2 (Item Scores/Right or Wrong) × *J* (Score levels/0 to *K*) tables. The 2 (Groups) × 2 (Item Scores) × *J* (Score levels) contingency table for each item can be viewed in 2 × 2 slices (See Table 2.3).

The null DIF hypothesis for the MH method is as follows:

$$H_0 : (A_j / B_j) / (C_j / D_j) = 1 \qquad j = 1, \ldots, K. \qquad (1)$$

Table 2.3 The 2 (Groups) × 2 (Item Scores) × J (Score levels)

Item Score

	Right (1)	Wrong (0)	Total
Focal	A_j	B_j	n_{Fj}
Reference	C_j	D_j	n_{Rj}
Total	m_{1j}	m_{0j}	T_j

The hypothesis means that the odds of getting the item correct at a given level of the total score is the same in both the focal group and the referenced group, across all J levels of the total score (matching variable).

Before discussing the MH procedure in detail, the ways of comparing proportions will be introduced. There are three ways of comparing proportions: difference of proportions, relative risk, and odds ratio. Table 2.4 displays notation for the distributions for a 2 × 2 contingency table.

Table 2.4 Notation for Joint, Conditional, and Marginal Probabilities The Distributions of a 2 × 2 Contingency Table

	Column		
Row	1	2	Total
1	p_{11} $(p_{1\|1})$	p_{12} $(p_{2\|1})$	p_{1+} (1.0)
2	p_{21} $(p_{1\|2})$	p_{22} $(p_{2\|2})$	p_{2+} (1.0)
Total	p_{+1}	p_{+2}	1.0

Difference of proportions.

For subjects in row 1, $p_{1|1}$ is the probability of response 1, and $(p_{1|1}, p_{2|1}) = (p_{1|1}, 1-p_{1|1})$ is the conditional distribution of the binary response. Two rows can be compared using the difference of proportions, $p_{1|2} - p_{1|1}$. Comparison on response 2 is equivalent to comparison on response 1, since

$$p_{2|2} - p_{2|1} = (1 - p_{1|2}) - (1 - p_{1|1}) = p_{1|1} - p_{1|2}.$$

The difference of proportions falls between -1.0 and +1.0. It equals zero when rows 2 and 1 have identical conditional distributions. The response is statistically independent of the row classification when $p_{1|1} - p_{1|2} = 0$. The null hypothesis in this case is that the population probabilities (π) are equal:

$$H_0 : \pi_{1|1} = \pi_{1|2}. \qquad (2)$$

Relative risk.

A difference in proportions of fixed size may have greater importance when both proportions are close to 0 or 1 than when they are near the middle of the range. In such a case, the ratio of proportions is also a useful descriptive measure. For 2 × 2 tables, the relative risk is the ratio:

$$p_{1|1} / p_{1|2}.$$

This ratio can be any nonnegative real number. A relative risk of 1.0 corresponds to independence. The null hypothesis in this case is the following:

$$H_0 : \pi_{1|1}/\pi_{1|2} = 1. \qquad (3)$$

Odds ratio.

Refer to Table 2.4. Within row 1, the odds that the response is in column 1 instead of column 2 is defined to be:

$$\Omega_1 = p_{1|1}/p_{2|1}.$$

Within row 2, the corresponding odds equals:

$$\Omega_2 = p_{1|2}/p_{2|2}.$$

The ratio of the odds Ω_1 and Ω_2, $\theta = \Omega_1/\Omega_2$, is called the *odds ratio*. From the definition of odds using joint probabilities,

$$\theta = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

When all cell probabilities are positive, independence of Row and Column is equivalent to $\theta = 1$. When $1 < \theta < \infty$, subjects in row 1 are more likely to make the first response than are subjects in row 2; that is, $p_{1|1} > p_{1|2}$.

The constant odds ratio hypothesis.

In their seminal paper, Mantel and Haenszel developed a chi-square test of the null DIF hypothesis against a particular alternative hypothesis known as the constant odds ratio hypothesis,

$$H_1 : (A_j / B_j) = \alpha\ (C_j / D_j) \qquad j = 1, \ldots, K \text{ and } \alpha \neq 1. \qquad (4)$$

Note that $\alpha = 1$ corresponds to the null hypothesis of (1). The parameter α is called the common odds ratio in the k 2×2 tables because under H_1, the value of α is the odds ratio:

$$\alpha = \frac{A_j/B_j}{C_j/D_j} = \frac{A_jD_j}{B_jC_j}$$

for all $j = 1, \ldots, K$.

The Mantel-Haenszel chi-square test statistic.

There is a chi-square test associated with the MH approach, namely a test of the null hypothesis, H_0: $\alpha_j = 1$,

$$MH\text{-}CHISQ = \frac{(|\Sigma_j A_j - \Sigma_j E(A_j)| - .5)^2}{\Sigma_j \ Var(A_j)}$$

where,

$$E(A_j) = n_{Rj}m_{1j}/T_j, \ \ \mathrm{VAR}(A_j) = n_{Rj}n_{Fj}m_{1j}m_{0j}/T_j^2(T_j - 1),$$

and where the -.5 in the expression for MH-CHISQ serves as a continuity correction to improve the accuracy of the chi-square percentage points as approximations to the observed significance levels. The quantity MH-CHISQ is distributed approximately as a chi-square with one degree of freedom.

Estimate of constant odds ratio.

Mantel-Haenszel (1959) provided an estimate of the constant odds ratio,

$$\alpha_{MH} = [\Sigma A_jD_j/T_j] \,/\, [\Sigma B_jC_j/T_j].$$

This is an estimate of DIF effect size which the early chi-square procedures could not produce. The metric estimate ranges from 0 to ∞ with a value of 1 playing the

roll of a null value of no DIF. The α_{MH} can be transformed by the natural logarithm to give:

$$\beta_{MH} = \log_e(\alpha_{MH}).$$

The β_{MH} is a signed index with a value of 0 indicating null DIF. A positive value signifies DIF in favor of the Reference group, and a negative value indicates DIF in favor of the Focal group.

MH DIF in item difficulty metrices.

The Educational Testing Service (ETS) worked with item difficulty estimates using the *delta metric* (Δ), which has a mean of 13 and a standard deviation of 4. Holland and Thayer (1988) proposed that:

$$\Delta_{MH} = -2.35 \ln(\alpha_{MH})$$

be used as a measure of the amount of DIF. Δ_{MH} has the interpretation of being a measure of DIF in the scale of differences in item difficulty as measured in the *delta metric* of ETS. Based on this new scale, ETS developed three categories which labeled A, B, and C (Zieky, 1993) to reflect the degree of DIF in test items.

The exact definitions of the categories follow:

Category A) $MH\chi^2$ test does not show significant differences from zero
OR
absolute value of Δ_{MH} is less than 1.0

Category B) $MH\chi^2$ test shows significant differences from zero and the absolute

value of Δ_{MH} is at least 1.0
AND EITHER
(1) absolute value of Δ_{MH} is less than 1.5
OR
(2) MHχ^2 test does not show significantly greater than 1.0

Category C) MHχ^2 test shows significantly greater than 1.0
AND
the absolute value of Δ_{MH} is 1.5 or more.

It appears that items falling into category B are examined for potential bias; however, items falling into category C are typically removed or replaced.

This MH procedure is currently one of the most popular procedures for detecting DIF because of its computational simplicity, ease of implementation, and associated test of statistical significance. In fact, the MH approach is the statistical test possessing the most statistical power for detecting departures from the null DIF hypothesis that are consistent with the constant odds ratio hypothesis. In other words, the MH-test is based on a normal approximation to the uniformly most powerful unbiased test of the null hypothesis, that the odds are equal to unity. However, there is an important assumption that α_j is uniform. In a comparison study, Rogers and Swaminathan (1993) showed that the MH-test is less sensitive to nonuniform DIF than the logistic regression procedure which will be discussed next.

Logistic Regression Procedure

Swaminathan and Rogers (1990) applied the logistic regression model to the detection of DIF. Uniform DIF exists if there is no interaction between ability level and group membership. It means that the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability. Nonuniform DIF exists if there is interaction between ability level and group membership. It means that the discrepancies in the probabilities of a right answer for the two groups is not consistent across all ability levels.

The logistic regression model which uses the examinee as the unit of analysis has the following form:

$$P(u = 1) = \frac{e^{z}}{(1 + e^{z})}$$

where $z = \delta + \tau_1\theta + \tau_2 g + \tau_3(\theta g)$. In this model, θ is the observed ability level of the examinee, and g represents group membership ($g = .5$ if examinee is a member of group 1 or $g = -.5$ if examinee is a member of group 2). θg is the product of the two independent variables, g and θ. The parameter τ_2 models the group difference in performance on the item, and τ_3 corresponds to the interaction between group and ability level. For example, an item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$, and non-uniform DIF if $\tau_3 \neq 0$ (whether $\tau_2 = 0$ or not).

Testing hypotheses with logistic regression.

The coefficients of the logistic regression model are estimated by the method of maximum likelihood (Agresti, 1990). The hypotheses of interest are that $\tau_2 = 0$ and $\tau_3 = 0$.

These two hypotheses can be tested simultaneously with the null hypothesis that $H_0: C\tau = 0$, where:

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The statistic for testing the joint hypothesis is:

$$\chi^2 = \tau' C' (C \Sigma C')^{-1} C \tau'$$

which has the χ^2 distribution with 2 degrees of freedom. When the test statistic exceeds $\chi^2_{\alpha;2}$, the null hypothesis ($H_0: C\tau = 0$) that there is no DIF is rejected.

Item Response Theory Approaches

Item Response Theory

The item response theory (IRT) is a mathematical function that relates the probability of a particular response on an item to overall examinee ability. This function is known as the item characteristic curve (ICC). Although an infinite number of IRT models are possible, only a few models are in current use. The three most popular uni-dimensional IRT models are the one-, two-, and three-parameter logistic models. These models are appropriate for dichotomous item

response data. The simplest IRT model is the following Rasch model (Rasch, 1980) which has one item parameter:

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

where:

$P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly

b_i is the item i difficulty parameter

n is the number of items in the test, and

e is a transcendental number whose value is 2.718.

The $P_i(\theta)$ is an S-shaped curve with values between 0 and 1 over the ability scale. The parameter for an item is the point on the ability scale where the probability of a correct response is 0.5. Figure 2.5 is a plot of what this function looks like for three items of different difficulty. Note that the ICCs for this model are parallel to one another. However, in many applications of IRT, it has been found that one does not get a good fit to the data with 1-PL. A common cause of misfit is that the ICCs of all items are not always parallel. The two parameter logistic model allows for different slopes. This parameter, usually denoted a, characterizes the slope of the item characteristic curve, and is often called the item's discrimination.

The 2-PL is:

$$P_i(\theta) = \frac{1}{1 + e^{-a(\theta - b_i)}} \qquad i = 1, 2, \ldots, n.$$

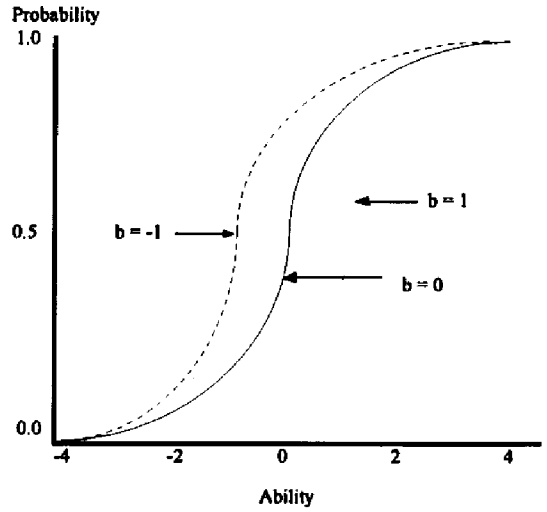

Figure 2.5 Item Characteristic Curves for 1-PL Model (Three levels of difficulty)

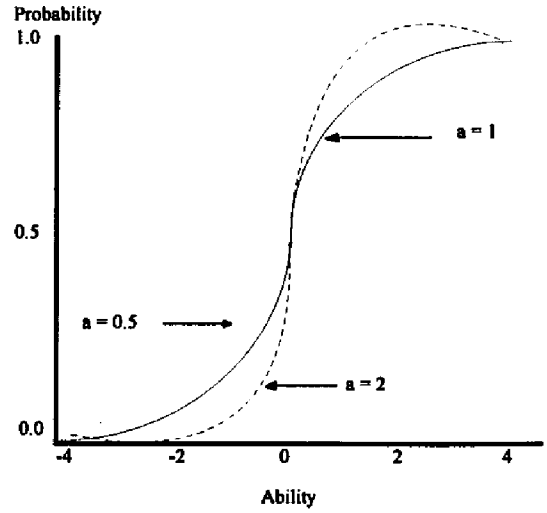

Figure 2.6 Item Characteristic Curves for the 2-PL model

Figure 2.6 shows a plot demonstrating the variation in slopes. Shown are items which have high discrimination ($a = 2$), average discrimination ($a = 1$), and low discrimination ($a = .5$).

A problem may arise in applying the one- and two-parameter logistic models to data obtained from multiple-choice or true-false items because these tests permit correct responses from guessing. The three-parameter logistic (3-PL) model allows for the guessing factor. The 3-PL model is:

$$P_i(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}} \qquad i = 1, 2, \ldots, n.$$

The additional parameter in the model, c, is called the pseudo-guessing parameter. The Figure 2.7 shows a typical ICC of the three-parameter logistic model.

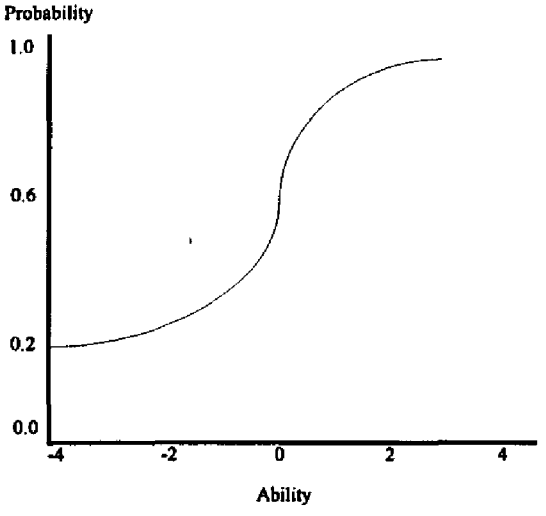

Figure 2.7 Item Characteristic Curve for the 3-PL Model

The fundamental assumptions in IRT are unidimensionality and local independence. The meaning of unidimensionality is that only one ability is measured by a set of items in a test. The local independence of items means that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent (See, Lord, 1980; Hambleton, Swaminathan, & Rogers, 1991).

Item Response Theory as Applied to DIF

The item characteristic curve (ICC) in the IRT models is a means for comparing the responses of two different groups to the same item. A difference between the two ICCs of two groups indicates that Group 1 and Group 2 examinees at the same ability level do not have same probability of success on the item. Camilli and Shepard (1994) technically defined that "DIF is said to occur whenever the conditional probability, $P(\theta)$, of a correct response differs for two groups (p. 58)". Because an ICC is determined by its a, b, and c parameters, it can be described as differences in the a, b, and c parameters when DIF is conceptualized as different ICCs for two groups. Figure 2.8 shows relative performance on a single item for individuals of Group 1 and Group 2. In Figure 2.8, the ICC for Group 1 is shifted to the left, indicating that this item is relatively easier for Group 1.

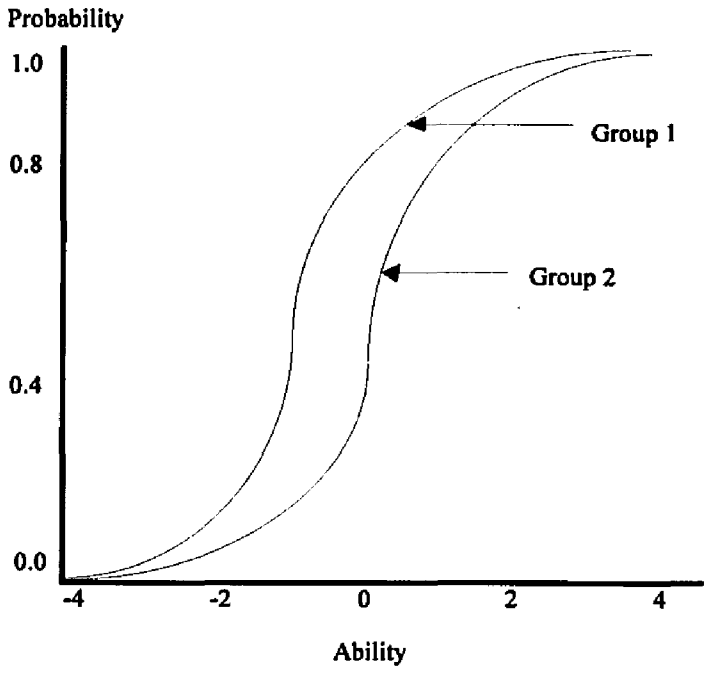

Figure 2.8 Relative Performance for Two Groups (Different b Parameters)

Figure 2.9 shows two ICCs which differ in all three parameters. In Figure 2.9, lower ability level students from Group 2 have higher probability of success although the item is more difficult overall for the group 2. As noted in the Mantel-Haenszel and logistic regression procedures, the latter case shows nonuniform DIF.

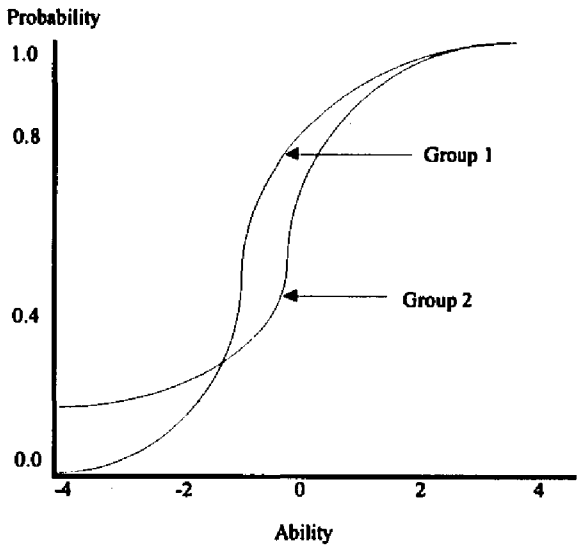

Figure 2.9 Relative Performance for Two Groups (Different a, b, and c parameters)

Scaling Item Parameters

Before the index of DIF is calculated, the estimates of the item parameters must be expressed on the same scale for each subpopulation. There are two major methods to obtain comparable item parameters for two groups in DIF analysis. One way to do this is to transform the estimates of the b_g's so that they have a mean of 0 and a standard deviation of 1 (Lord, 1980). An alternative procedure involves standardizing on θ, the ability or latent trait scores. Camilli and Shepard (1994) labeled this method the *separate sample method.* Because the scales for the ability (latent trait score) estimates are different for the two

groups, the scales for the estimates of difficulty and discrimination are also different. One group is placed on the scale of the other by a scale transformation that requires two constants expressed in terms of the slope (β) and the intercept (α) of the theoretical linear relationship between the *b*s of Group 1 and Group 2. The equations used in the transformation follow:

$$b^* = \alpha + \beta b$$

$$a^* = a/\beta,$$

where a and b are the discrimination and difficulty parameter estimates for a particular item for Group 1, and a^* and b^* are parameter estimates for the same item for Group 2. The procedure works for the one-, two-, and three-parameter logistic models. It should be noted that the guessing parameter, *c*, is not influenced by a change of scale and there is no need to equate the *c*s for the two groups.

The other method, also labeled by Camilli and Shepard (1994), is the *anchor test method.* In this method, the test items are divided into two sets: the *studied* item set, and the *anchor* item set. The studied item set has only one item and the anchor item set has the remaining items on the test. During the estimation, the item parameters for the anchor items are constrained to be identical for the two groups, but the parameters for the studied item are allowed to vary. The scales for the two sets of studied item parameters are equivalent

because they are linked through the equality constraint on the anchor items' parameters.

Comparing ICCs

There are two approaches to examining DIF with IRT models. First is the method to measure the size or magnitude of DIF. Second is the method to test the statistical significance of DIF.

Test for the magnitude of DIF.

Rudner (1977) has proposed calculating the area between the ICCs for two groups by using the formula:

$$A_g(signed) = \sum_{\theta=-4}^{\theta=4} .005[P_1g(\theta) - P_2g(\theta)]$$

where $P_{1g}(\theta)$ and $P_{2g}(\theta)$ refer to the value of the ICC for each of the two groups. The value of $P_{ig}(\theta)$ is calculated for each value of θ from -4 to 4 in steps of .005. The probability of a correct response for Group 2 is subtracted from that of Group 1. Therefore, if Group 1 performed better, the index will be positive. However, if ICCs cross as in Figure 2.9 (nonuniform case), the following formula can be used:

$$A_g(unsigned) = \sum_{\theta=-4}^{\theta=4} .005|P_1g(\theta) - P_2g(\theta)|$$

Since this equation converts negative values of $P_{1g}(\theta) - P_{2g}(\theta)$ to positive values, the positive and negative differences do not cancel each other to any extent.

Linn and Harnisch (1981) pointed out that Rudner's simple area method is flawed because "a simple comparison of item characteristic curves sometimes suggests differences in cases where there may be relatively few observations for one of the groups being compared (p. 115)." They proposed that the area be weighted to reflect the distribution of estimated θ's. Shepard and Others (1984) proposed "self-weighing" sums-of-squares indices in which squared differences between the ICCs for two groups are summed across the obtained estimates of θ only. Likewise, Camilli and Shepard (1994) presented two probability difference measures of DIF. The measures were labeled as "*signed probability difference controlling for θ (SPD-θ)* "and "*unsigned probability difference controlling for θ (UPD-θ)*". They maintained that "if the purpose of DIF analysis is to inspect test items that may be biased against Focal group examinees, measures of DIF should emphasize ICC differences in the range of θ where most Focal group examinees score (p. 67)." The SPD-θ and UPD-θ measures are as follows:

$$SPD\text{-}\theta = \frac{\sum_{j=1}^{nF} \Delta P_j}{nF} \qquad UPD\text{-}\theta = \frac{\sum_{j=1}^{nF} (\Delta P_j)^2}{nF}$$

where $\Delta P_j = P_R(\theta_j) - P_F(\theta_j)$.

Test for the statistical significance of DIF.

Lord (1980) proposed a test of the null hypothesis that $b_{1g} = b_{2g}$ and $a_{1g}=a_{2g}$. To test the significance of the b difference ($H_0 : \Delta b = 0$), the standard errors of b_{1g} and b_{2g} should computed. The standard error of the difference ($\Delta b = b_{1g} - b_{2g}$) follows:

$$S_{\Delta_b} = \sqrt[2]{S_1^2 + S_2^2}$$

A statistic for testing the difference for significance (Lord, 1980) can be given

$$t = \frac{\Delta b}{S_{\Delta_b}} ,$$

where t has approximately a normal distribution. A separate significance test can be made of the null hypothesis that $H_0 : \Delta a = 0$. However, Lord (1980) also proposed a chi-square test to test both those hypotheses simultaneously (See Lord, 1980, p. 223).

Thissen and Others (1993) proposed four methods of testing significance of DIF. The methods are implemented by comparing the relative fit of two models, a compact model and an augmented model which were labeled by Judd and McClelland (1989). The goal of the comparison is to see whether the additional parameters in the augmented model are significantly different from zero. A simpler model with a single ICC for the two groups is always preferable

to a more complex model in which each group has its own ICC. As noted in the discussion of IRT, the null hypothesis of no DIF is that there are no significant differences between the item parameter(s) for the two groups. Thus, to test for item *i*, the ML (Maximum likelihood) estimates of the parameters of the compact model (with no DIF for item *i*) and the likelihood under that model should be computed, and the ML estimates and likelihood of the augmented model should be computed by some parameters representing differences between the item *i* parameters for the two groups. The likelihood ratio statistic provides a test of the significance of DIF on *k* degrees of freedom, in which *k* is the number of item parameters which differ between the two groups. Thissen and Others (1993) discussed General IRT, Loglinear IRT, Limited Information IRT, and IRT as implemented in item drift techniques, IRT-D^2 (See pp. 71-85).

Standardization Approach

With a desire to avoid contamination caused by model misfit, Dorans and Kulick (1983) proposed an IRT-like approach standardization method that compared empirical item response curves in which a total score was used as an estimate of ability. According to the standardization method, "an item is exhibiting DIF if the *expected performance* on an item differs for examinees of equal ability from different groups" (Dorans & Holland, 1993; pp. 43-44). The expected performance on an item can be operationalized by nonparametric item

test regressions. If there is a difference in an empirical item test regression, it indicates DIF.

One of the important principles of the standardization method is to use all available appropriate data to estimate the conditional item performance of each group at each level of the matching variable. Let $E_{1g}(I|M)$ define the empirical item test regression for Group 1, and let $E_{2g}(I|M)$ define the empirical item test regression for group 2, where I is the item score variable and M is the matching variable. If $E_{1g}(I|M) \neq E_{2g}(I|M)$, it indicates DIF. Dorans and Kulick (1986) presented two item discrepancy indices, *the standardized P-difference and the root-mean-weighted squared difference* which follow:

$$D_{STD} = \sum_{s=1}^{S} K_s \, [P_1 s - P_2 s] \; / \; \sum_{s=1}^{S} K_s \quad ,$$

$$RMWSD = \sqrt{\sum_{s=1}^{S} K_s \, (P_1 s - P_2 s)^2 \; / \; \sum_{s=1}^{S} K_s} \quad ,$$

where $[K_s/\Sigma K_s]$ is the weighing factor at score level s supplied by the standardization group to weight differences in performance between the two groups (P_{1s} and P_{2s}). As discussed in IRT approaches, the D_{STD} is a signed index and the RMWSD is an unsigned index.

DIF Assessing Methods for Performance Tasks

Logistic Discriminant Function Analysis

As noted in Chapter 1, recently three procedures to detect DIF for performance assessment were introduced. Miller and Spray (1993) proposed the *logistic discriminant function analysis* for DIF identification of polytomously scored items. They first introduced three extensions of the logistic regression procedure to accommodate polytomous items. However, they pointed out that the methods require assumptions that may not be warranted in a DIF analysis. For example, the proportional odds models require the *equal-slopes regression lines assumption.* They insisted that the logistic discrimination function analysis procedure overcomes the problems.

The logistic discriminant function analysis is a transformed model of the logistic regression model which has been used successfully to detect DIF in simulations of dichotomous item responses (Swaminathan & Rogers, 1990). The logistic regression model can be written as:

$$Prob(U|X,G) = \frac{e^{(1\cdot U)(\beta 0+\beta 1 X+\beta 2 G+\beta 3 X\cdot G)}}{1+e^{(\beta 0+\beta 1 X+\beta 2 G+\beta 3 X\cdot G)}}$$

where U is each dichotomous item response, X is a test score, and G is a group indicator variable. In this model, the item response variable U is treated as a random variable and X and G are fixed, explanatory variables. Miller and Spray

(1993) developed the logistic regression procedure to estimate Prob($G|X,U$) when G is fixed and U is random. This procedure is more appropriate than normal discriminant function analysis because the normality of the explanatory variables is often violated (See Bull & Donner, 1987; Demaris, 1992, pp.61-70).

The logistic discriminant function analysis model to DIF detection in polytomous item responses can be written as:

$$Prob(G|X,U) = \frac{e^{(1-G)(-\alpha 0-\alpha 1X-\alpha 2U-\alpha 3XU)}}{1 + e^{(-\alpha 0-\alpha 1X-\alpha 2U-\alpha 3XU)}}$$

In this model (Say Model 1), the response variable U need not be restricted to only two categories but can take on any one of the J values associated with each item.

Significance test with logistic discriminant function analysis.

The coefficients in the logistic regression model are estimated by the method of maximum likelihood. Therefore, hypotheses based on comparisons of different models can be tested by likelihood ratio statistics in the same manner as the logistic regression models.

Specifically, the significance of α_3 is tested by first fitting the hierarchical model given by:

$$Prob(G|X,U) = \frac{e^{(1-G)(-\alpha 0-\alpha 1X-\alpha 2U)}}{1 + e^{(-\alpha 0-\alpha 1X-\alpha 2U)}} \qquad (Model\ \ 2)$$

The difference in the log of the likelihood functions obtained from Model 1 and Model 2 is used to test for the significance of α_3. This is the test for nonuniform DIF. The significance of α_2 is tested by next fitting the null model, given by:

$$Prob(G|X,U) = \frac{e^{(1-G)(-\alpha 0-\alpha 1X)}}{1 + e^{(-\alpha 0-\alpha 1X)}} \qquad (Model \quad 3).$$

The difference in the log of the likelihood function obtained from Model 2 and Model 3 is used to test for the significance of α_2, which presents a test for uniform DIF. Miller and Spray (1993) presented simultaneous confidence bands of the Scheffe type for items with significant DIF (See pp.110-111, 119-121).

Extensions of Mantel-Haenszel Procedure

Zwick and Others (1993) introduced two extensions of Mantel-Haenszel statistic for polytomous items.

Mantel approach for ordered response categories.

Zwick and Others (1993) introduced a one degree-of-freedom test of conditional association for the case of ordered response categories, *the Mantel approach for ordered response categories*. In order to apply the method in the DIF context, first index numbers to the response categories are assigned and second, compare the item means for members of the two groups who have been matched on a measure of proficiency. The data are organized into a $2 \times C \times K$ contingency table, where C is the number of response categories and K is the

number of levels of the matching variable. Table 2.5 shows a $2 \times C$ contingency table as one of the K levels.

Table 2.5. Data of a 2 × C Table

	Item Score				
Group	y_1	y_2	y_3 ...	y_C	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k} ...	n_{RCk}	N_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k} ...	n_{FCk}	N_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k} ...	n_{+Ck}	N_{++k}

The statistic proposed by Mantel, reformulated by Zwick and Others (1993), is:

$$Mantel\ \chi^2 = \frac{[\Sigma\ F_k - \Sigma\ E\ (Fk)]^2}{\Sigma\ Var\ (F_k)}$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable. Under H_0, the Mantel χ^2 has a chi-square distribution with one degree of freedom. In DIF applications, rejection of H_0 indicates that members of the two groups who are similar in overall proficiency tend to differ in their mean performance on the studied item.

GMH statistic for nominal data.

In the notation of Table 2.5,

$$A'_k = (n_{R1k}, n_{R2k}, n_{R3k}, \ldots, n_{R(C-1)k}),$$

$$E(A'_k) = n_{R+k}n'_k/n_{++k},$$

$$n'_k = (n_{+1k}, n_{+2k}, n_{+3k}, \ldots, n_{+(C-1)k}),$$

$$V(A_k) = n_{R+k}n_{F+k}\,[\{n_{++k}\mathrm{diag}(n_k) - n_k\, n'_k\}/\{n^2_{++k}(n_{++k} - 1)\}],$$

and $\mathrm{diag}(n_k)$ is a $(C - 1)(C - 1)$ diagonal matrix with element n_k.

The test statistic is:

$$\text{GMH } \chi^2 = [\,\Sigma\, A_k - \Sigma\, E\,(A_k)]'[\Sigma\, V(A_k)]^{-1}[\Sigma\, A_k - \Sigma E\,(A_k)].$$

This statistic has a chi-square distribution with C-1 degrees of freedom under the null hypothesis of no conditional association between group membership and response. The GMH statistic does not explicitly take into account the possible ordering of response categories. However, Zwick and Others argued that it provides for the comparison of the two groups in terms of their entire response distribution.

Combined *t*-test Procedures

Welch and Hoover (1993) proposed two combined *t*-test methods for use in detecting DIF in polytomously scored items. Both methods test a hypothesis about the difference between two mean values of the two subgroups of interest. The mean values are summed across k-independent tests. However, the second method, HW3 labeled by Welch and Hoover, uses a weighting procedure for different sample sizes at each level of ability between two subgroups.

A Combined *t*-test method for unweighted sample size.

The first method requires an assumption of homogeneity of variances of performance test scores at each of the *k*-ability levels for the two subgroups.

Using a total test score as an external criterion or matching variable, all possible score categories on the external criterion are used as the ability levels. For each ability level, a separate *t*-test is conducted. The equation used for each level can be written as:

$$t = \frac{\overline{X_F} - \overline{X_R}}{\sqrt{\frac{(S_F^2 n_F + S_R^2 n_R)}{n_F + n_R - 2}\left(\frac{1}{n_F} + \frac{1}{n_R}\right)}}$$

where $\overline{X_F}$ is the mean of students in the Focal group, $\overline{X_R}$ is the mean of students in the Reference group, S_F^2 is the variance of students in the Focal group, S_R^2 is the variance of students in the Reference group, n_F is the number of students in the Focal group, and n_R is the number of students in the Reference group.

Winer (1971) described two statistics for combining several independent tests on the same hypothesis. One was χ^2 statistic for the chi-square distribution and the other was z statistic for the normal distribution. The z statistic is applicable for a test of significance for combining the results of several independent *t* tests. The test equation can be written as:

$$z = \frac{\Sigma t_j}{\sqrt{\Sigma\left(\frac{df_j}{df_j - 2}\right)}}$$

where df_j represents the degrees of freedom associated with t_j, and z is normally distributed ($N(0,1)$).

A combined t-test method for weighted sample size.

As noted earlier most tests of DIF involve subgroups with different sample sizes at each of the k-ability levels. Welch and Hoover (1993) proposed another statistic, labeled HW3 by them, with a weighting procedure to more accurately represent the size of the Focal and Reference groups. The test equation can be written as:

$$z = \frac{\left(\frac{d_1/S_1^2 + \ldots\ldots + d_k/S_k^2}{1/S_1^2 + \ldots\ldots + 1/S_k^2} \right)}{1/\sqrt{1/S_1^2 + \ldots\ldots + 1/S_k^2}}$$

where z is normally distributed ($N(0,1)$). For the equation, the numerator is the weighted average of the k-independent effect sizes and the denominator is the standard error of the weighted average. In the equation, $d_1 \ldots d_k$ are k-independent effect size estimates and their standard errors are $S_1 \ldots S_k$. The unbiased effect size can be computed by the following equation:

$$d = \left(1 - \frac{3}{4(n_R + n_F - 2) - 1} \right) \times \sqrt{\frac{n_R + n_F}{n_R n_F}} \times t$$

In this method, the standard errors of the individual effect size estimators provide weights for optimally combining effect sizes across tests (Welch & Hoover, 1993).

Summary

A number of methods for detecting DIF were reviewed in this chapter. Although IRT models provide a powerful and accurate method for the study of DIF, the three IRT models have different advantages and disadvantages in practical use. The one PL model has a limitation in generalization, while the three PL model requires larger sample sizes for the efficient estimation of item and ability parameters (Thissen, Steinberg, & Wainer, 1993). Moreover, proper implementation and interpretation of IRT methods require considerable sophistication. That is why the MH procedure and the logistic regression analysis have gained wide acceptance as useful methods.

As discussed in Chapter 1, although DIF procedures for dichotomous items are well established, there is not yet a clear concept about estimates of DIF in performance assessment. One major problem, the difficulty in identifying a matching variable, arises when developing a DIF analysis strategy. A fundamental problem that arises is that an entire performance assessment may consist of very few tasks (e.g., a single item in a written composition test). Therefore, defining an appropriate matching variable for performance tasks is

less than straightforward. One possible strategy is to match subjects using an external criterion to the performance assessment, such as the score on multiple-choice items (Zwick et al., 1993). The three methods studied in the Monte Carlo study for performance tasks used the total test scores as a matching variable that consisted of both dichotomous and polytomous items. These studies have relevance because many current applications of performance assessments incorporate both conventional and performance items. However, some believe that the future direction for performance is likely be pure performance assessment. Therefore, there is a need to study DIF in performance assessment using a matching variable which consists of only polytomously scored items.

Also, based on previous research findings (Donoghue, Holland, & Thayer, 1993; Miller & Spray, 1993; Zwick, Donoghue, & Grima, 1993; Welch & Hoover, 1993), the factors that will be varied in this study are as follows: mean abilities of Focal and Reference groups, type of DIF, number of DIF items in the matching variable, sample size, and ratio of Focal and Reference group sizes. In the next chapter, a description of the research methods will be presented.

CHAPTER THREE: METHODOLOGY

Overview

This chapter presents a description of the research methods employed in this study. The first part describes the design of the simulation. In the design of the simulation, several factors that impact the performance of the differential item functioning (DIF) procedures are discussed. Because these factors are related with the research questions in this study, the presentation goes along with the research questions. Based on these factors, the general strategy of the simulation is described with the flowchart of the procedure.

Following this, operational definitions of the conditions that are to be manipulated are presented. The operational definitions for the factors which may be cruxes of this study (i.e., uniform and nonuniform DIF, matching variable) clarify the design of this study and also assist the understanding of the interpretation of the results in Chapter 4.

Then, a technical description of the procedures of the study are presented. Finally, the statistical analyses are described.

Simulation Design

As noted in Chapter 1, the research questions were developed based on findings from previous research. As a consequence, the design of the present simulation was guided by the research questions. As noted above, this literature

has shown several factors which impact the performance of the DIF procedures considered in this study. These are: (1) mean ability of Focal and Reference groups, (2) the size of the sample, (3) ratio of Focal group sample size to Reference group sample size, (4) the type of DIF, and (5) the proportion of DIF items on the matching variable.

The following values of the parameters were selected based on the results of previous studies:

- Research Question 1 (Which statistical method is the most powerful when there are differences in ability between Reference and Focal groups?) The values of *Mean ability* (A) for the populations of Focal and Reference groups were based on two conditions: even ($A_F \sim N(0,1)$, $A_R \sim N(0,1)$), one standard deviation in favor of Reference group ($A_F \sim N(-1,1)$, $A_R \sim N(0,1)$), where A_F is the writing ability of Focal group, A_R is the writing ability of Reference group, and N is normal distribution function.
- Reseach Question 2 (Which statistical method is the most powerful when the size of the sample is relatively small?)

  The value of *Size* of the sample varied as follows: 1500, 500.
- Research Question 3 (Which statistical method is the most powerful when there are unequal sample sizes between Reference and Focal groups?) The value of *Ratio* of Reference group sample size to Focal

group sample size varied as follows: 1:1, 2:1.

- Research Question 4 (Which statistical method is the most powerful when nonuniform DIF exists?)

The value of *Type* of DIF varied as follows: Uniform, Nonuniform.

- Research Questions 7 and 8 (Is there any effect of the proportion of DIF in a test on detecting DIF? If any effect of the proportion of DIF items in a test exists on detecting DIF, which statistical method is the most efficient for this condition?)

The values of *Proportion* of DIF items on the matching variable varied as follows: 10%, 20%, 30%, 40%.

The factors considered above define two (Mean ability) populations and eight (Size × Ratio × Type) experimental setups (i.e., 16 simulations), and another eight (Type × Proportion) experimental setups (i.e., eight simulations). Therefore, a total of twenty-four simulations were accomplished in this study. For each experimental setup and population, fifty replications were made. In each instance, sampling occurred with replacement.

It should be noted that the methodology for Research Question 5 (Which statistical method demonstrates consistent control of Type I error under the null hypothesis?) was designed when the first item in 10 simulated items was generated as a non-DIF item (i.e., the first item represents the null hypothesis).

This item parameter is presented in Table 3.1. Also, regarding Research Question 6 (Which statistical method is the most powerful for detecting DIF across all conditions?), it might be considered that the values of parameters for all conditions were designed for this question.

The overall strategy of the simulation was accomplished in six steps. In step 1, ten hypothetical polytomously scored items, each with a different degree or magnitude of DIF, were generated and fixed for all simulated conditions. In step 2, a variety of conditions related to the factors discussed above were generated. In step 3, based on a chosen condition, random samples from the populations for Reference and Focal groups were generated. In step 4, given a sample value of ability and the fixed item parameters of the ten test items, a response to a given item was calculated through the partial credit model. In step 5, data from the responses were used to compute the statistics for the three procedures. In step 6, steps 3 through 5 were repeated 50 times and the number of times the method rejected an item at the .05 level was recorded. The flowchart of the procedure for each condition is as follows:

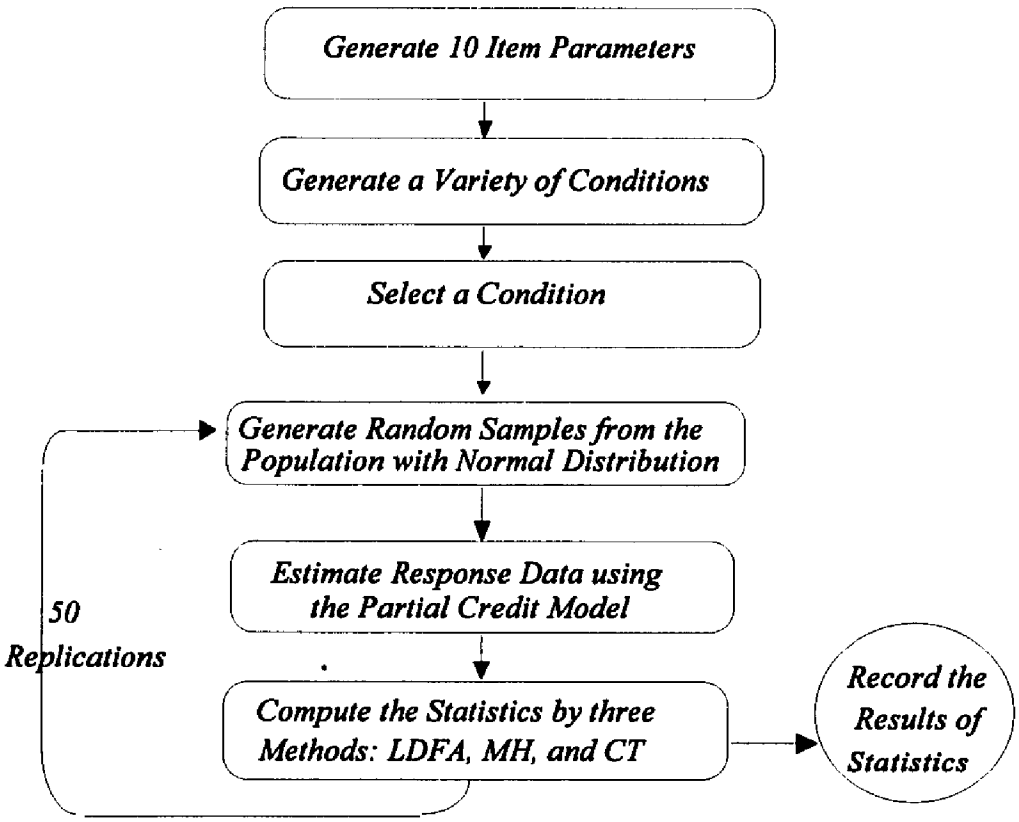

Figure 3.1 Flowchart of the Simulation Procedures

Operational Definition and Design of the Factors

In this section, operational definitions and design of the conditions which are to be manipulated are presented.

Uniform and Nonuniform DIF

In this study, uniform DIF is operationally defined as DIF in which the Item Characteristic Curves (ICCs) for two groups are different and do not cross, while nonuniform DIF is operationally defined as DIF in which the ICCs for two

groups are different but cross at some point on the theta (θ) scale. The DIF for and against a group cancel each other to some degree. Positive and negative DIF may cancel each other entirely, even though it is extremely rare in practice (Camilli & Shepard, 1994). The condition under which positive and negative DIF completely cancel each other, ideal for the purpose of examining the relative empirical power of the three procedures, was designed for the nonuniform DIF in this study. When a student's ability was equal or greater than the mean of the ability distribution (i.e., $\theta \geq 0$), the DIF item was in favor of the Focal group, while if a student's ability was less than the mean value (i.e., $\theta < 0$), the DIF item was favor of the Reference group.

Matching Variable

As discussed in Chapters 1 and 2, all three procedures employed in this study use a total test score as the matching variable. It is assumed that the total test score is unbiased. However, if a biased item exists in a test, the total test score cannot be considered free from bias. In order to examine the relative statistical powers for three procedures based on the assumption of DIF theory, it is necessary to use an index which is unbiased. Therefore, the integer transformed theta (θ) score (ITT) for each individual student was developed and used as a matching variable in this study. The formula for estimating the ITT is the following:

$$ITT = INT\ [(theta \times 10) + 50]$$

where INT is the function of integers. For the second research question of this study, a total test score which includes the scores of all possible DIF items in the test was used as a matching variable. Therefore, the matching variable is operationally defined as the ITT scores or total test scores.

Technical Description of the Procedure

The details of the simulation are described below.

Step 1. Polytomous Item Parameters

As noted above, item parameters were fixed for 10 simulated polytomous items with 10 varying degrees of differentiation between the Reference and Focal groups. The parameters of these 10 items, which originated from the design of Welch and Hoover (1993), are presented in Table 3.1.

The parameters assume that each item contains three steps as defined by the partial credit model (Master & Wright, 1984). According to the partial credit model, the probability of obtaining a score of x on item i for an examinee with proficiency θ is given by:

$$P_{ix}(\theta) = \frac{\exp\sum_{j=1}^{x}(\theta - d_{ij})}{1+\sum_{k=1}^{m}\exp\sum_{j=1}^{k}(\theta - d_{ij})}, \quad x = 1,2,..,m,$$

Table 3.1 Performance Test Item Parameter Estimated for R and F Groups

Item	Step	Reference	Focal
1	1	-1.75	-1.75
	2	-.25	-.25
	3	.25	.25
2	1	-1.75	-1.70
	2	-.25	-.20
	3	.25	.30
3	1	-1.75	-1.65
	2	-.25	-.15
	3	.25	.35
4	1	-1.75	-1.60
	2	-.25	-.10
	3	.25	.40
5	1	-1.75	-1.55
	2	-.25	-.05
	3	.25	.45
6	1	-1.75	-1.50
	2	-.25	.00
	3	.25	.50
7	1	-1.75	-1.45
	2	-.25	.05
	3	.25	.55
8	1	-1.75	-1.40
	2	-.25	.10
	3	.25	.60
9	1	-1.75	-1.35
	2	-.25	.15
	3	.25	.65
10	1	-1.75	-1.30
	2	-.25	.20
	3	.25	.70

where d_{ij} represents the difficulty of making the transition from category m to m+1. Based on this model, a three-step partial credit model can be used to describe a performance item with possible score values of 1 to 4. Step 1 represents the transition from Category 1 to Category 2. Step 2 represents the transition from Category 2 to Category 3. Step 3 represents the transition from Category 3 to Category 4. For a three step item ($m = 3$), three separate operating curves are defined. The first curve shows the probability of scoring 2 rather than 1 on the item. In each of the operating curves in a test, the curve for the second step in Item 1 is to the right of the curve for the first, so that the second step is considered to be more difficult than the first step (Welch & Hoover, 1993). Master (1988) stated that:

> At any estimated level of competence β_n, the partial credit model provides the widths P_{ni0}, P_{ni1}, P_{ni2} of the four outcome regions at that level. These widths can be interpreted either as the estimated probabilities of a student at that level of competence responding in outcome categories, 0, 1, 2, and 3, or as the expected proportions of students at that level of competence responding in these four categories (p.294).

Step 2. Response generation

Random samples of simulated examinees (size N_F and N_R) were generated so that actual testing behavior was simulated between the Reference and Focal groups. One set of conditions was simulated based on uniform DIF, while the second set of conditions was simulated based on nonuniform DIF. Based on the

2 populations (mean ability) and 4 experimental setups (sample size × ratio of two groups), one set of conditions of this simulation study for each type of DIF had eight conditions. These conditions were:

Condition A $A_F \sim N(0, 1), N_F=1,500$
$A_B \sim N(0,1), N_B=1,500$

Condition B $A_F \sim N(-1, 1), N_F=1,500$
$A_B \sim N(0,1), N_B=1,500$

Condition C $A_F \sim N(0, 1), N_F=750$
$A_B \sim N(0,1), N_B=1,500$

Condition D $A_F \sim N(-1, 1), N_F=750$
$A_B \sim N(0,1), N_B=1,500$

Condition E $A_F \sim N(0, 1), N_F=500$
$A_B \sim N(0,1), N_B=500$

Condition F $A_F \sim N(-1, 1), N_F=500$
$A_B \sim N(0,1), N_B=500$

Condition G $A_F \sim N(0, 1), N_F=250$
$A_B \sim N(0,1), N_B=500$

Condition H $A_F \sim N(-1, 1), N_F=250$
$A_B \sim N(0,1), N_B=500$

For the second research question of this study, a total test score which includes the scores of all possible DIF items in the test was used as a matching variable. In order to examine the effect of the proportion of DIF items in a test on detecting DIF by three procedures, four conditions were varied in this study.

The four conditions are the following:

Condition I:

One item (10%) is a DIF item with a magnitude of 0.25.

Condition II:

Two items (20%) are DIF items with magnitudes of 0.15 and 0.35.

Condition III:

Three items (30%) are DIF items with magnitudes of 0.15, 0.25, and 0.35.

Condition IV:

Four items (40%) are DIF items with magnitudes of 0.05, 0.15. 0.35. and 0.45.

The average size of the DIF magnitude for all conditions was controlled as 0.25. Thus, it was assumed that the overall effect size of the DIF items on the total test score was directly proportional to the number of DIF items in a test.

Step 3. Item Responses

Given a sample value of ability (A) and the fixed item parameters of the test items, probabilities of responding (P(A)) for the items were calculated by the partial credit model and compared against three thresholds -- P_1(A), P_2(A), and P_3(A) -- between 0 and 1 to generate the performance item data. A response to a given item, Y_{ij} was made according to:

Y_{ij} = 1, [0.00 < U < P_1(A)]

2, [P_1(A) < U < P_2(A)]

3, $[P_2(A) < U < P_3(A)]$

4, $[P_3(A) < U < 1.00]$

Step 4. Computation of item statistics

For each given ability level (A), the test item responses were then available. This information was used to compute the statistics defined as Logistic Discriminant Function Analysis, the Extension of Mantel-Haenszel, and the Combined t-test Procedure for each of the 10 polytomous items.

Step 5. Replications

Steps 1 through 4 were repeated 50 times for each set of the 24 conditions.

Step 6. Summary of performance by three methods

Each statistic was converted to a probability value to examine the number of times the method rejected an item at the .05 level. Each procedure's performance was also summarized by examining the distribution of values across the 50 replications.

Statistical Analysis

Logistic Discriminant Function Analysis

Regression coefficients were estimated as described in Chapter 2 for three models: (1) the full model, which includes the matching variable (total score), item, and total score-by-item interaction; (2) the reduced model, which includes only the total score and item terms; and (3) the null model, which includes only

the total score. For each model, a likelihood ratio chi-squared goodness-of-fit statistic, G^2, was computed. Differences in G^2 (G^2_{diff}) provided a test of significance of the improvement in fit by adding an additional term. The statistic of G^2_{diff} was obtained from $G^2_{diff} = -2 (\ln L (i+1) - \ln L (i))$, where $L (i + 1)$ and $L (i)$ are the likelihoods for adjacent hierarchical models. Therefore, the statistics of G^2_{diff} between the full and reduced models tested for nonuniform DIF. The statistics of G^2_{diff} between the reduced and null models tested for uniform DIF. Significant values for nonuniform DIF, uniform DIF, or both were determined.

The Extension of Mantel-Haenszel Procedure

From two extensions of the Mantel-Haenszel statistic described in Chapter 2, *Mantel approach* for ordered response categories was used in this study. First, index numbers to the response categories were assigned, and second, the item means for members of the two groups that were matched on a measure of proficiency were compared. The data were organized into a $2 \times C \times K$ contingency table, where C is the number of response categories and K is the number of levels of the matching variable. Table 3.2 shows a $2 \times C$ contingency table as one of the K levels.

Table 3.2. Data of a 2 × C Table

	Item Score				
Group	y_1	y_2	y_3 ...	y_C	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k} ...	n_{RCk}	N_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k} ...	n_{FCk}	N_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k} ...	n_{+Ck}	N_{++k}

A summary statistic, with one degree of freedom, was computed as:

$$Mantel\ \chi^2 = \frac{[\Sigma\ F_k - \Sigma\ E\ (Fk)]^2}{\Sigma\ Var\ (F_k)}$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable. F_k is defined as:

$$F_k = \sum_c y_c n_{Fck}$$

Rejection of H_0 indicated that members of the two groups who are similar in overall proficiency tend to differ in their mean performance on the studied item.

The Combined t-test Procedure

From the two combined *t*-test methods described earlier, the combined *t*-test method using non-weighting procedure (HW1) was employed in this study. Using the matching variable (total score), all possible score categories on the total score were used as levels of ability. At each level of ability a separate *t*-test was conducted. Mean performance was computed for the two groups at each of

the *k*-ability levels, and a separate between-groups *t* statistic at each score level was computed. Finally, a test of significance for combining the results of a series of *t* tests was performed. As noted in the extension of MH procedure, if the null hypothesis (H_0: z = 0) is rejected, members of the two groups who are similar in overall proficiency differ in their mean on the studied item.

CHAPTER FOUR: RESULTS AND DISCUSSION

Overview

This chapter will be divided into two main sections. The first section will present the results of the simulation study which was performed with the matching variable which is completely free from bias [i.e., the integer transformed theta (ITT) score]. This section will be separately discussed based on six different research questions. The second part of this chapter will present results of the simulation study which used the total test score as the matching variable. This second section is related to the last two research questions.

As noted in Chapter 3, one way to assess the performance of the statistics is to identify the number of times, across the 50 replications, an item is flagged at a probability level of .05. This allows for a comparison of empirical power functions for the three procedures.

Category I : Simulations Based on the Integer Transformed Theta Scores

Research Question 1: Which statistical method is the most powerful when there are differences in ability between Reference and Focal groups?

The results of the four testing simulations based on unequal ability conditions, each replicated 50 times, are summarized below. As presented in Table 3.1, Item 1 represents the null hypothesis (i.e., unbiased item).

Table 4.1 Percentage of Flagged Items at .05 Level by Statistic for Conditions B and D (Unequal Ability)

	Condition B (N_F = 1,500; N_B = 1,500)			Condition D (N_F = 750; N_B = 1,500)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	6	4	10	6	6	6
2	12	14	6	10	10	8
3	46	46	40	28	26	20
4	82	84	68	76	80	60
5	100	100	94	84	84	76
6	100	100	100	100	100	98
7	100	100	100	100	100	100
8	100	100	100	100	100	100
9	100	100	100	100	100	100
10	100	100	100	100	100	100

Table 4.2 Percentage of Flagged Items at .05 Level by Statistic for Conditions F and H (Unequal Ability)

	Condition F (N_F = 500; N_B = 500)			Condition H (N_F = 250; N_B = 500)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	2	2	6	2	0	0
2	6	12	8	8	10	2
3	22	28	16	18	16	12
4	48	48	38	28	24	14
5	56	56	38	28	34	28
6	92	90	76	66	70	54
7	92	90	74	78	78	66
8	98	96	92	92	90	82
9	100	100	100	92	92	84
10	100	100	98	96	98	98

Items 2 through 10 represented increasing degrees of falsity or departure from the null hypothesis (H_0). Table 4.1 through 4.2 summarizes the percentage of flagged items for each of the procedures at the .05 level.

When unequal ability distributions for the Reference and Focal groups existed, several patterns were present in the identification of items. For Condition B, Table 4.1 shows that the logistic discriminant function analysis (LDFA) and the Mantel-Haenszel (MH) followed a fairly consistent pattern in detecting items. For flagged Items 2, 3, 4, and 5, LDFA and MH procedures exhibited slightly higher empirical power than the combined *t*-test (CT) procedure in Conditions B and D. When sample sizes decreased to 500:500 (i.e., Condition F in Table 4.2), LDFA and MH methods showed apparently higher statistical power than CT. For Condition F, MH flagged Items 2 and 3 at a somewhat higher rate than LDFA.

The performance of all procedures was affected by the decrease in sample sizes (i.e., comparison of Conditions B and D with F and H). Also, the ratio of sample sizes (i.e., 1:1 and 1:2) seemed to affect the performance of the procedures. The ratio of sample size (i.e., 1:2) had a stronger impact for smaller samples than for larger samples (i.e., comparison of the differences between Conditions B and D with the differences between Conditions F and H).

Table 4.3 provides a summary of the empirical power functions presented in Tables 4.1 and 4.2. Table 4.3 presents the overall average percentage of the performance of three methods on detecting DIF items.

Table 4.3 Average Percentage for Flagged Items at .05 level by LDFA, MH, and CT (Based on Unequal Ability Conditions)

Condition	LDFA	MH	CT
B (1500:1500)	82.2	82.7	78.7
D (750:1500)	77.6	77.8	73.6
F (500: 500)	68.2	68.9	60.0
H (250: 500)	56.2	56.9	48.9
Total	71.1	71.6	65.3

LDFA and MH show a similar number for the average percentage. However, MH exhibited slightly higher power than LDFA across all four conditions. Note that CT's average percentage number is 65.3 which is approximately 6 percent lower than those of LDFA and MH.

Table 4.4 provides pairwise comparisons of the performance of LDFA, MH, and CT. Nine comparisons (Item 2 through 10) were available for each pair of procedures. Item 1, the no DIF item, was not included when summing across items. The numbers in Table 4.4 represent the number of times a given procedure identified more items than the other method of the pair. The tie columns indicate the number of times the two procedures identified an identical number of times.

Table 4.4 Pairwise Comparisons of the Performance of LDFA, MH, and CT (Based on Unequal Ability Conditions)

Condition									
	LD	MH	Tie	LD	CT	Tie	MH	CT	Tie
B	0	2	7	4	0	5	4	0	5
D	1	1	7	5	0	4	5	0	4
F	3	2	4	7	1	1	8	0	1
H	3	4	2	7	1	1	8	0	1
Total	7	9	20	23	2	11	25	0	11

As Table 4.4 indicates, LDFA and MH procedures outperformed the CT method when considering the total across all four conditions. Although the pairwise comparison of LDFA and MH showed a preference for MH, the difference in performance between LDFA and MH was very small (i.e., 7:9 with 20 tie).

In conclusion, the results indicate that LDFA and MH appeared to outperform CT when there were differences in ability between Reference and Focal groups.

Research Question 2: Which statistical method is the most powerful when the size of the sample is relatively small?

When the size of the sample for the Reference and Focal groups is relative small (i.e., $N_F = 500$; $N_B = 500$ or $N_F = 250$; $N_B = 500$), several patterns were

present in the identification of items. Tables 4.5 and 4.6 summarize the percentage of flagged items for each of the procedures at the .05 level.

Tables 4.5 and 4.6 show that LDFA and MH exhibit a fairly consistent pattern in identifying DIF items. For Condition E (equal ability case), LDFA and MH flagged Items 3, 4, 5, and 6 at a higher rate than CT. When unequal ability distributions for the Reference and Focal groups existed (i.e., Condition F), LDFA and MH procedures exhibited slightly higher empirical power than the CT for flagged Items 3, 4, 5, 6, and 7.

Table 4.5 Percentage of Flagged Items at .05 Level by Statistic for Conditions E and F ($N_F = 500$; $N_R = 500$)

	Condition E (Equal Ability)			Condition F (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	6	6	4	2	2	6
2	6	8	8	6	12	8
3	38	30	20	22	28	16
4	36	32	26	48	48	38
5	82	78	70	56	56	38
6	90	90	78	92	90	76
7	92	94	94	92	90	74
8	100	100	96	98	96	92
9	100	100	100	100	100	100
10	100	100	100	100	100	98

Table 4.6 Percentage of Flagged Items at .05 Level by Statistic for Conditions G and H ($N_F = 250$; $N_R = 500$)

	Condition G (Equal Ability)			Condition H (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	14	12	6	2	0	0
2	2	6	4	8	10	2
3	16	16	16	18	16	12
4	40	36	28	28	24	14
5	54	58	50	28	34	28
6	72	72	60	66	70	54
7	96	90	76	78	78	66
8	94	94	86	92	90	82
9	98	98	96	92	92	84
10	100	100	100	96	98	98

When the ratio of sample sizes changed to 1:2 (i.e., Condition G), LDFA and MH flagged Items 3, 4, 5, 6, 7, and 8 at a somewhat higher rate than CT. When unequal ability distributions for the Reference and Focal groups existed and the ratio of sample sizes is 1:2 (Condition H), LDFA and MH methods showed slightly higher empirical power than CT for flagged Items 2, 3, 4, 5, 6, 7, 8, and 9. These patterns are similar to those in Research Question 1.

Table 4.7 provides a summary of the empirical power functions presented in Tables 4.5 and 4.6. It presents the overall average percentage of the performance of the three procedures on detecting DIF items.

Table 4.7 Average Percentage for Flagged Items at .05 level by LDFA, MH, and CT (Based on Small Size of Samples)

Condition	LDFA	MH	CT
E (500: 500)(Equal Ability)	71.6	70.2	65.8
F (500: 500)(Unequal Ability)	68.2	68.9	60.0
G (250: 500)(Equal Ability)	63.6	63.3	57.3
H (250: 500)(Unequal Ability)	56.2	56.9	48.9
Total	64.9	64.8	58.0

LDFA and MH show almost same number (64.9 and 64.8) of the average percentage. However, LDFA exhibited a slightly higher empirical power than MH for equal ability cases (Conditions E and G), while MH showed a somewhat higher statistical power than LDFA for unequal ability cases (Conditions F and H). Note that CT's average percentage number is 58.0 which is about 7 percent lower than those of LDFA and MH.

Table 4.8 provides pairwise comparisons of the performance of LDFA, MH, and CT. It should be noted that nine comparisons (Item 2 through 10) were available for each pair of statistics. Item 1, the no DIF item, was not included when summing across items.

Table 4.8 Pairwise Comparisons of the Performance of LDFA, MH, and CT (Based on Small Size of Samples)

Condition									
	LD	MH	Tie	LD	CT	Tie	MH	CT	Tie
E	3	2	4	5	2	2	5	0	4
F	3	2	4	7	1	1	8	0	1
G	2	2	5	6	1	2	7	0	2
H	3	4	2	7	1	1	8	0	1
Total	11	10	15	25	5	6	28	0	8

As Table 4.8 indicates, LDFA and MH procedures outperformed the CT method when considering the total across all four conditions. Although the pairwise comparison of LDFA and MH appeared almost the same (i.e., 11:10 with 15 ties), their comparisons against CT indicated a preference for MH. Note that the comparison of LDFA and CT was 25:5 with 6 ties, while the comparison of MH and CT was 28:0 with 8 ties.

In conclusion, the results indicate that LDFA and MH appear to outperform CT when small sample sizes of Reference and Focal groups exist.

Research Question 3: Which statistical method is the most powerful when there are unequal sample sizes between Reference and Focal groups?

The results of the four testing simulations based on unequal sample size conditions, each replicated 50 times, are summarized in Tables 4.9 and 4.10.

Table 4.9 Percentage of Flagged Items at .05 Level by Statistic for Conditions C and D (N_F = 750; N_R = 1,500)

	Condition C (Equal Ability)			Condition D (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	6	4	6	6	6	6
2	16	16	12	10	10	8
3	38	44	22	28	26	20
4	68	66	54	76	80	60
5	98	98	88	84	84	76
6	100	100	98	100	100	98
7	100	100	100	100	100	100
8	100	100	100	100	100	100
9	100	100	100	100	100	100
10	100	100	100	100	100	100

Table 4.10 Percentage of Flagged Items at .05 Level by Statistic for Conditions G and H (N_F = 250; N_R = 500)

	Condition G (Equal Ability)			Condition H (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	14	12	6	2	0	0
2	2	6	4	8	10	2
3	16	16	16	18	16	12
4	40	36	28	28	24	14
5	54	58	50	28	34	28
6	72	72	60	66	70	54
7	96	90	76	78	78	66
8	94	94	86	92	90	82
9	98	98	96	92	92	84
10	100	100	100	96	98	98

When unequal sample sizes between the Reference and Focal groups existed, several patterns were present in the identification of items. For

Conditions C and D, Table 4.9 shows that LDFA and MH followed a fairly consistent pattern in detecting items. For flagged Items 2, 3, 4, and 5, LDFA and MH procedures exhibited slightly higher empirical power than CT procedure. When sample sizes decreased to 250:500 (i.e., Conditions G and H in Table 4.10), LDFA and MH methods showed apparently higher statistical power than CT. As discussed in the results for Research Question 1, the factor of unequal sample sizes between Reference and Focal groups had a stronger impact for smaller samples than for larger samples (i.e., comparison of the Conditions Cand D with the Conditions G and H).

Table 4.11 provides a summary of the empirical power functions presented in Tables 4.9 and 4.10. Table 4.11 presents the overall average percentage of the performance of three methods for detecting DIF items.

Table 4.11 Average Percentage for Flagged Items at .05 level by LDFA, MH, and CT (Based on Unequal Sample Sizes)

Condition	LDFA	MH	CT
C (750:1500)(Equal Ability)	80.0	80.4	74.9
D (750:1500)(Unequal Ability)	77.6	77.8	73.6
G (250: 500)(Equal Ability)	63.6	63.3	57.3
H (250: 500)(Unequal Ability)	56.2	56.9	48.9
Total	69.4	69.6	63.7

LDFA and MH show a similar number for the average percentage. Note that CT's average percentage number is 63.7, which is approximate 6 percent lower than those of LDFA and MH.

Table 4.12 provides pairwise comparisons of the performance of LDFA, MH, and CT. Nine comparisons (Item 2 through 10) were available for each pair of statistics.

Table 4.12 Pairwise Comparisons of the Performance of LDFA, MH, and CT (Based on Unequal Sample Size)

Condition	LD	MH	Tie	LD	CT	Tie	MH	CT	Tie
C	1	1	7	5	0	4	5	0	4
D	1	1	7	5	0	4	5	0	4
G	2	2	5	6	1	2	7	0	2
H	3	4	2	7	1	1	8	0	1
Total	7	8	21	23	2	11	25	0	11

As Table 4.12 indicates, LDFA and MH procedures outperformed the CT method when considering the total across all four conditions. Although the pairwise comparison of LDFA and MH appeared almost the same (i.e., 7:8 with 21 ties), their comparisons against CT indicated a preference for MH. Note that the comparison of LDFA and CT was 23:2 with 11 ties, while the comparison of MH and CT was 25:0 with 11 ties.

In conclusion, the results indicate that LDFA and MH appear to outperform CT when there are unequal sample sizes between Reference and Focal groups.

Research Question 4: Which statistical method is the most powerful when nonuniform DIF exists?

The results of the eight testing simulations based on nonuniform DIF cases are summarized in Table 4.13 through 4.16.

Table 4.13 Percentage of Flagged Items at .05 Level by Statistic for Conditions A and B (N_F = 1,500; N_B = 1,500)(Nonuniform)

	Condition A (Equal Ability)			Condition B (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	4	4	0	4	4	4
2	4	6	2	6	6	4
3	16	4	2	16	12	2
4	26	2	2	44	30	8
5	58	4	2	58	34	12
6	76	4	2	84	54	24
7	88	14	10	88	56	18
8	98	10	8	98	82	46
9	100	10	8	100	88	50
10	100	8	4	100	96	68

When equal ability distributions for the Reference and Focal groups existed (Condition A), only LDFA performed successfully to detect DIF items. However, when there were unequal ability distributions between the Reference and Focal groups (Condition B), MH and CT exhibited much higher empirical power than when equal ability distributions existed. Note that this pattern consistently appeared across all conditions (See Conditions C, D, E, F, G, and H).

Table 4.14 Percentage of Flagged Items at .05 Level by Statistic for Conditions C and D (N_F = 750; N_R = 1,500)(Nonuniform)

	Condition C (Equal Ability)			Condition D (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	8	6	6	6	8	10
2	14	6	4	10	8	8
3	18	6	8	14	12	10
4	26	4	0	26	14	6
5	52	10	2	64	58	24
6	62	10	4	82	66	20
7	90	8	6	80	66	38
8	84	2	2	90	80	58
9	98	14	8	98	96	58
10	100	8	4	100	88	48

Table 4.15 Percentage of Flagged Items at .05 Level by Statistic for Conditions E and F (N_F = 500; N_R = 500)(Nonuniform)

	Condition E (Equal Ability)			Condition F (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	8	6	8	4	4	4
2	2	0	2	6	6	6
3	6	2	8	16	10	8
4	16	4	4	22	12	8
5	30	8	10	42	30	18
6	42	2	4	54	34	24
7	60	8	8	54	38	18
8	70	4	6	74	54	24
9	84	8	2	78	42	22
10	84	8	8	86	56	26

Table 4.16 Percentage of Flagged Items at .05 Level by Statistic for Conditions G and H ($N_F = 250$; $N_R = 500$)(Nonuniform)

	Condition G (Equal Ability)			Condition H (Unequal Ability)		
Item	LDFA	MH	CT	LDFA	MH	CT
1	2	2	4	2	2	2
2	0	2	2	2	2	6
3	8	6	4	4	6	4
4	12	4	10	14	8	6
5	12	4	2	22	14	8
6	32	10	4	32	22	16
7	38	8	6	48	30	22
8	48	12	6	60	32	16
9	50	2	2	66	50	22
10	62	4	6	72	50	22

As presented in Tables 4.13 through 4.16, LDFA dramatically outperformed MH and CT. The average percentages of MH and CT for detecting DIF items on Condition A were only 6.8 and 4.4, respectively. Note that the average percentages (.068 and .044) yield almost the same value of Type I error (.05) in this study.

Like the results of the uniform DIF analysis, the performance of LDFA was also influenced by sample sizes and the ratio of sample sizes between the Reference and Focal groups. It is interesting to note that the LDFA procedure exhibited slightly higher empirical power in unequal ability distribution cases (i.e., Conditions B, D, F, and H) than in equal ability distribution cases (i.e.,

Conditions A, C, E, and G). However, the result was reversed in the uniform DIF case because the performances of all three procedures exhibited higher empirical power in equal ability distribution cases than in unequal ability distribution cases.

Table 4.17 provides a summary of empirical power functions presented in Tables 4.13 through 4.16.

Table 4.17 Average Percentage for Flagged Items at .05 level by LDFA, MH, and CT (Based on Nonuniform DIF Conditions)

Condition	LDFA	MH	CT
A (1500:1500)(Equal Ability)	62.9	6.8	4.4
B (1500:1500)(Unequal Ability)	66.0	50.9	25.8
C (750:1500)(Equal Ability)	60.4	7.6	4.2
D (750:1500)(Unequal Ability)	62.7	54.2	30.0
E (500: 500)(Equal Ability)	43.8	4.9	5.8
F (500: 500)(Unequal Ability)	48.0	31.3	17.1
G (250: 500)(Equal Ability)	29.1	5.8	4.7
H (250: 500)(Unequal Ability)	35.6	23.8	13.6
Total	51.1	23.2	13.2

Note that the MH and CT procedures show much higher empirical power in unequal ability distribution cases than in equal ability distribution cases. As Table 4.17 indicates, the MH and CT procedures appear to detect DIF items by chance when there is equal ability (note the average percentages of the procedures in conditions A, C, E, and G). However, MH and CT exhibited somewhat desirable power for detecting DIF in the unequal ability distribution

cases. In particular, MH showed empirical power close to that of LDFA in unequal ability distribution cases.

Table 4.18 provides pairwise comparisons of the performance of LDFA, MH, and CT. Although MH shows its empirical power close to that of LDFA in unequal ability distribution cases, Table 4.18 indicates that only LDFA is well suited for identifying nonuniform DIF. In addition, MH exhibited higher empirical power than CT for detecting nonuniform DIF.

Table 4.18 Pairwise Comparisons of the Performance of LDFA, MH, and CT (Based on Nonuniform DIF Conditions)

Condition									
	LD	MH	Tie	LD	CT	Tie	MH	CT	Tie
A	8	1	0	9	0	0	6	0	3
B	8	1	0	9	0	0	9	0	0
C	9	0	0	9	0	0	7	1	1
D	9	0	0	9	0	0	8	0	1
E	9	0	0	7	1	1	1	5	3
F	8	0	1	8	0	1	8	0	1
G	8	1	0	8	0	1	5	2	2
H	7	1	1	7	1	1	8	1	0
Total	66	4	2	66	2	4	52	9	11

In conclusion, LDAF is the most powerful statistics to detect DIF items for performance assessment when there is nonuniform DIF.

Research Question 5: Which statistical method demonstrates consistent control of Type I error under the null hypothesis?

Table 4.19 Type I error rates by Three Procedures

Condition	Uniform			Nonuniform		
	LDFA	MH	CT	LDFA	MH	CT
A	6	6	4	4	4	0
B	6	4	10	4	4	4
C	6	4	6	8	6	6
D	6	6	6	6	8	10
E	6	6	4	8	6	8
F	2	2	6	4	4	4
G	14	12	6	2	2	4
H	2	2	0	2	2	2
Total	48	40	42	38	36	38
Mean	6%	5%	5.25%	4.75%	4.5%	4.75%

Table 4.19 shows Type I error rates of three procedures based on all eight conditions. Table 4.19 indicates that MH demonstrated the most consistent control of Type I error under the null hypothesis. For uniform DIF, MH and CT showed more consistent control of Type I error than LDFA. However, when nonuniform DIF existed, all three procedures demonstrated almost the same rate of Type I error. In fact, it seemed that there was no significant difference of Type I error rate between the three methods.

Research Question 6: Which statistical method is the most powerful for detecting DIF across all conditions?

Regarding Research Question 6, the LDFA seems the most powerful statistic for detecting DIF across all conditions. As discussed in Research Questions 1 through 3, although LDFA and MH followed a fairly consistent

pattern in detecting items in uniform DIF conditions, LDFA showed much higher empirical power than MH in nonuniform conditions. Thus, the results indicated a preference for LDFA. Specifically, LDFA demonstrated the strongest power to detect nonuniform DIF for polytomously scored items.

Although MH appears to have slightly higher empirical power to control Type I error and shows a similar level of power to LDFA when unequal ability exists between two groups, LDFA demonstrates the highest empirical power for identifying DIF across all conditions.

Category II : Simulations Based on the Total Test Scores

For the purposes of this study, the results related to Research Questions 7 and 8 will be discussed together.

Research Question 7: Is there any effect of the proportion of DIF in a test on detecting DIF?

Research Question 8: If any effect of the proportion of DIF items in a test exists on detecting DIF, which statistical method is the most efficient for this condition?

Based on the results from a preliminary simulation using ITT scores as a matching variable, Condition A ($A_F \sim N(0, 1)$, N_F=1,500; $A_B \sim N(0,1)$, N_B=1,500) was used for uniform DIF, while Condition B ($A_F \sim N(-1, 1)$, N_F=1,500; $A_B \sim N(0,1)$, N_B=1,500) was used for nonuniform DIF. Note that Condition A was the best condition for detecting uniform DIF items, while Condition B was the most

effective condition for identifying nonuniform DIF items. In other words, the two conditions maximized the empirical powers of three methods-- LDFA, MH, and CT.

In order to examine effects of the proportion of DIF items in a test, four conditions based on differences in the proportion of DIF presented in Chapter 3 were analyzed for both uniform and nonuniform DIF cases. The results of the eight simulations are summarized in Tables 4.20 through 4.27.

Table 4.20 Percentage of Flagged Items at .05 Level by Statistic when One DIF Item exists (N_F = 1500; N_R = 1500)(Uniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	14	14	6
2	0	8	8	4
3	0	8	6	4
4	0	8	10	4
5	0	4	8	4
6	0	6	6	0
7	0	12	14	6
8	0	6	8	6
9	0	4	4	8
10	0.25	100	100	100
Type I error		0.078	0.087	0.047

Table 4.21 Percentage of Flagged Items at .05 Level by Statistic when Two DIF Items exist (N_F = 1500; N_R = 1500)(Uniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	20	18	14
2	0	24	24	16
3	0	20	20	24
4	0	16	16	6
5	0	22	24	26
6	0	20	18	14
7	0	26	26	22
8	0	14	16	12
9	0.15	56	60	52
10	0.35	100	100	100
Type I error		0.2025	0.2025	0.1675

Uniform DIF

When ten percent of the test items were DIF items, all three procedures exhibited identical empirical power for detecting items with a DIF of 0.25 magnitude. However, LDFA, MH, and CT had a Type I error of 0.078, 0.087, and 0.047, respectively. LDFA and MH had a higher Type I error rate than CT.

When twenty percent of the test items were DIF items, all three procedures also exhibited identical performance (100%) on detecting items with a DIF of 0.35 magnitude. However, the three procedures showed somewhat different empirical powers for identifying items with a DIF of 0.15 magnitude. Note that the empirical power rates of three methods for detecting items with a

Table 4.22 Percentage of Flagged Items at .05 Level by Statistic when Three DIF Items exist (N_F = 1500; N_B = 1500)(Uniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	32	32	36
2	0	40	40	36
3	0	40	40	38
4	0	40	38	30
5	0	40	40	32
6	0	48	46	40
7	0	38	38	24
8	0.15	44	42	32
9	0.25	98	98	92
10	0.35	100	100	100
Type I error		0.397	0.391	0.337

Table 4.23 Percentage of Flagged Items at .05 Level by Statistic when Four DIF Items exist (N_F = 1500; N_B = 1500)(Uniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	54	52	46
2	0	64	64	46
3	0	60	66	50
4	0	60	58	50
5	0	56	56	46
6	0	68	68	58
7	0.05	12	10	14
8	0.15	26	26	16
9	0.35	100	100	100
10	0.45	100	100	100
Type I error		0.603	0.607	0.493

DIF of 0.15 magnitude are still higher than their Type I error rates. This means that the probability for detecting items with a DIF of 0.15 magnitude is still higher than the probability for labeling items as DIF when they are not.

When thirty percent of the test items were DIF items, the empirical power rates of the three methods for detecting DIF items of 0.15 magnitude were almost the same as their Type I error rates. Finally, when forty percent of the test items were DIF items, the empirical power rates of the three methods for detecting the DIF items of 0.15 magnitude are much smaller than their Type I error rates.

Now, the probability for detecting items with a DIF of 0.15 magnitude is lower than the probability for labeling items as DIF when they are not. Thus, when total scores are used as the matching criterion, a purified matching criterion seems to need to be develop through deleting DIF items in a test.

Nonuniform DIF

The result of the performances of the three procedures for nonuniform DIF was quite different from that of the uniform DIF. From the result of the eight simulations based on ITT scores for nonuniform DIF, the fact that LDFA outperformed MH and CT on identifying nonuniform DIF items was determined. Thus, the analysis of this section focused on the performance of LDFA.

Table 4.24 Percentage of Flagged Items at .05 Level by Statistic when One DIF Item exists (N_F = 1500; N_R = 1500)(Nonuniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	10	12	6
2	0	2	4	8
3	0	2	0	2
4	0	6	6	4
5	0	2	4	2
6	0	10	10	8
7	0	8	8	4
8	0	4	4	4
9	0	12	12	10
10	0.25	72	42	30
Type I error		0.062	0.067	0.053

Table 4.25 Percentage of Flagged Items at .05 Level by Statistic when Two DIF Items exist (N_F = 1500; N_R = 1500)(Nonuniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	8	10	6
2	0	4	6	4
3	0	10	12	4
4	0	8	8	8
5	0	10	10	6
6	0	8	8	6
7	0	16	16	10
8	0	8	8	0
9	0.15	50	26	12
10	0.35	96	72	58
Type I error		0.090	0.098	0.055

When ten percent of the items in a test were DIF items, LDFA exhibited a 72 percent empirical power for detecting items with a DIF of 0.25 magnitude. This means that the probability for identifying items with a DIF of 0.25 as a DIF item is 72 percent. When twenty percent of the items in a test were DIF items, LDFA exhibited a 50 percent empirical power for detecting items with a DIF of 0.15 magnitude, and a 96 percent statistical power for identifying items with a DIF of 0.35 magnitude.

Table 4.26 Percentage of Flagged Items at .05 Level by Statistic when Three DIF Items exist (N_F = 1500; N_B = 1500)(Nonuniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	6	8	2
2	0	14	12	8
3	0	20	20	14
4	0	14	18	8
5	0	10	10	4
6	0	6	6	4
7	0	10	12	6
8	0.15	36	14	8
9	0.25	84	44	24
10	0.35	96	56	40
Type I error		0.114	0.123	0.066

Table 4.27 Percentage of Flagged Items at .05 Level by Statistic when Four DIF Items exist (N_F = 1500; N_R = 1500)(Nonuniform)

Item	DIF Magnitude	LDFA	MH	CT
1	0	12	12	6
2	0	8	8	10
3	0	10	12	8
4	0	14	16	10
5	0	16	16	14
6	0	18	24	20
7	0.05	20	10	10
8	0.15	44	8	8
9	0.35	92	62	34
10	0.45	100	88	68
Type I error		0.130	0.147	0.113

When thirty percent of the items in a test were DIF items, LDFA showed a 36 percent power for identifying items with a DIF of 0.15 magnitude, an 84 percent statistical power for detecting items with a DIF of 0.25 magnitude, and a 96 percent empirical power for indentifying items with a DIF of 0.35 magnitude. Note that the empirical power rate of LDFA for identifying items with a DIF of 0.15 magnitude is still much higher than their Type I error rates (i.e., 36% vs. 11%). However, this was not true in the uniform case.

When forty percent of the items in a test were DIF items, the empirical power rate of LDFA for detecting items with a DIF of 0.15 magnitude is still much higher than its Type I error rate (i.e., 44% vs. 13%). This means that the probability for detecting items with a DIF of 0.15 magnitude is higher than the

probability for labeling *no* DIF items as DIF. In addition, the empirical power for detecting items with a DIF of 0.05 magnitude (20%) is slightly higher than its Type I error rate (13%).

Note that the rate of Type I error (i.e., labeling items as DIF when they are not) slowly increases from 6%, to 9%, 11% and 13%. However, the rate of Type I error rapidly increases from 8%, to 20%, 40% and 60% in uniform cases. Thus, the effect of the numbers of DIF items on the matching variable for identifying DIF items seems stronger for uniform DIF than for nonuniform DIF.

Regarding Research Question 7, some effects of the proportion of DIF items on the matching variable for identifying DIF in polytomously scored items were found: as mentioned above, the effects appeared stronger on uniform DIF than on nonuniform DIF.

Considering Research Question 8, LDFA still showed the highest statistical power to detect DIF when there were some DIF items on the matching variable. For uniform DIF, LDFA and MH exhibited almost the same empirical power, while LDFA showed apparent higher empirical power than MH to identify nonuniform DIF. CT demonstrated the most consistent control for Type I error, but the result was not surprising since CT was the least powerful statistic across all conditions. The results showed that CT exhibited the highest Type II error rate (i.e., failing to identify items which are DIF) across all conditions.

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview

Today, tests are widely used to select, promote, certify, and determine the competency of individuals within education and industry. As the use of testing in important decision-making has increased, and legal challenges to the uses of tests have become common, the issue of fairness in testing has become a major concern in the assessment of test validity.

The recent attention given to performance assessment raises the question of what is the effect of such assessments on various subgroups. The belief has been expressed that performance assessment is more valid than conventional assessment because it uses more direct (authentic) measures. However, some forms of performance assessment may be more likely than conventional assessment to produce construct-irrelevant factors which are a significant threat to validity. Therefore, as more attention is given to performance assessment, it becomes increasingly important to investigate the fairness of this testing. DIF analysis is one major component for monitoring fairness in testing.

This study was designed to evaluate the relative statistical powers of the three proposed procedures in identifying performance assessment items that function differentially for two subgroups of examinees. In the Monte Carlo study

a variety of conditions were generated and employed to examine the empirical powers of the procedures. This study dealt with two main concerns. The first concern was to determine which procedure was the most powerful for identifying DIF in performance assessment. The second concern was to determine the effect of the number of DIF items on the matching variable for detecting DIF. A summary of the findings and conclusions is contained in the following sections.

Summary of the Findings

First Issue: Determining the Most Powerful Procedure for DIF Analysis in Performance Assessment

Because Research Questions 1, 2, 3, 4 and 5 can be summarized in Research Question 6, the findings will be mainly discussed based on Research Question 6. The results of the total twenty-four simulations, each replicated 50 times, indicate a preference for the logistic discriminant function analysis (LDFA) procedure. LDFA appears to be more flexible than other available procedures for DIF detection in polytomously scored items.

In particular, the generalized Mantel-Haenszel (MH) and the Combined *t*-test (CT) procedures are not well suited for identifying nonuniform DIF. However, concerning Research Question 1 through 5, the MH method is not behind the LDFA in demonstrating its empirical power to detect uniform DIF.

Also, the MH procedure exhibited the highest empirical power to control Type I error even though the size of the differences of Type I error between the three methods was relatively small. The MH method became quite popular in the late 1980s because of its simplicity, intuitive appeal, and promise (Hambleton et al., 1993).

One repeated criticism of the MH procedure is that it is not useful for identifying nonuniform DIF (Swaminathan & Rogers, 1990; Miller & Spray, 1993). To overcome the major shortcoming of the MH procedure, Mazor, Clauser and Hambleton (1994) recommended using a variation on the MH statistic for detecting nonuniform DIF in dichotomously scored items. Unfortunately, its applicability and implementation in performance assessment has not been explored.

Many scholars (Hambleton et al., 1993; Camilli & Shepard, 1994) suggested using multiple methods for identifying DIF. Multiple methods can address the instability problem which undermines the utility of current methods and can address the shortcomings found in particular methods (Hambleton et al., 1993). However, if multiple procedures are used, Type I error (i.e., labelling items as DIF when they are not) can be easily inflated (Camilli & Shepard, 1994). Therefore, when two techniques (e.g., Methods A and B) show a consistent

pattern for detecting DIF items in most conditions, and Method A outperforms Method B in a specific condition, it is suggested to use Method A.

Through examining the results of the simulations of this study, it was observed that the LDFA procedure performed a consistent pattern with the MH method for identifying uniform DIF. In addition, the LDFA statistic outperformed the MH method for detecting nonuniform DIF. As discussed in Chapters 2 and 3, the analysis and interpretation of nonuniform DIF in the LDFA statistic seemed to be easy. In this case, why is the MH method preferred, causing a burden on the variation analysis of the statistic?

Second Issue: The Effect of the Proportion of DIF Items on the Matching Criterion

Concerning the effect of the number of DIF items on the matching variable, the effect of the proportion of biased items in the matching variable seemed significant for identifying DIF in performance assessment. The effect was stronger for identifying uniform DIF than for detecting nonuniform DIF.

As discussed earlier, with all three methods proposed in this study, the test score is used to match the Reference and Focal groups prior to comparing item performances. If the test score is not purified with potentially biased test items which would be eliminated, Type I error resulted from a contamination of the matching criterion increases. Note that the LDFA still exhibited a desirable

statistical power to detect DIF when there were some DIF items on the matching variable. More detailed discussion related to the issue of purified matching criterion will be presented in the conclusion.

Regarding the factors (i.e., conditions) implemented in this study, additional findings are as follows:

1. Across eight conditions, the LDFA and MH procedures demonstrated higher empirical power to detect uniform DIF in polytomously scored items than the CT procedure. The observation of Welch and Hoover (1993) was not supported by this result.

2. When differences in ability distributions existed, LDFA and MH also appeared to outperform the CT across all eight conditions for uniform DIF.

3. Unequal sample sizes between Reference and Focal groups influenced the empirical power more when unequal ability distributions between two groups existed. This result supported the observation of Welch and Hoover (1993).

4. For nonuniform DIF, the LDFA exhibited higher empirical power when the ability distributions were not equal.

5. For nonuniform DIF, MH and CT were not well suited; however, MH and CT procedures demonstrated some statistical power when the ability distributions were not equal.

6. For uniform DIF, the influence of the number of the DIF items in the matching variable appeared to be serious. The result of the study of Donoghue et al. (1993) using dichotomously scored items does not agree with the result of this study using polytomously scored items.

7. For nonuniform DIF, the LDFA still appeared promising when the effect of the DIF items in the matching variable existed.

Conclusions

Based on the findings from this study, it appears that three preliminary conclusions can be drawn:

1. For DIF analysis in performance assessments, the LDFA can be recommended as the preferred method to test constructors or practitioners. However, as Camilli wrote (1993), " the term DIF does not necessarily imply bias. It is a measure of an effect, and does not suggest the cause (p. 408)." Therefore, the logical analysis that comes after DIF analysis can also be recommended to test constructors or practitioners. As discussed in Chapter 1, both statistical DIF analysis and judgmental review are needed as checks on each other.

2. Through the use of the LDFA procedure for identifying DIF in performance assessment, the appropriateness of test use for different subgroups will be enlarged. As indicated through Tables 1.1 through 1.2 in Chapter One,

there is a great difference in academic performance between the minority and majority groups. If we assume that the difference is partially or entirely due to the bias in testing, the use of the appropriate method will ultimately assist in reducing this difference.

However, Camilli (1993) argued that DIF analyses address the problem of construct representation that concerns the internal validity of a test. It should be noted that as a means of demonstrating test fairness, the use of DIF analyses is by no means straightforward. A systematic bias in test scores cannot be detected because DIF methods are only sufficient for demonstrating the relative strengths of groups of examinees.

As mentioned by Angoff (1982), "These methods are, after all, only item-discrepancy methods; they should not be credited with a higher function than they are capable of serving (p.114)." The studies of DIF should not be treated as if they address broader questions of test bias and fairness. As argued in Chapter 1, DIF analysis is only one component of the extensive research that is needed to establish the validity and fairness of testing. Therefore, we should not claim that a test is unbiased and fair because it has been subjected to DIF analyses.

3. The effects of the number of DIF items on the matching variable seem significant for identifying DIF in performance assessment. If the matching

variable is itself biased to some degree, then the application of a DIF analysis will certainly be flawed. Thus, in order to decrease or minimize the effects of the proportion of DIF items on the matching variable, it is recommended to emphasize or enforce the judgmental analysis-- referred to as "sensitivity reviews" (Dorans, 1989)-- to evaluate biased items in a test before entering DIF analysis.

However, we know it is almost impossible to create a completely purified matching criterion (i.e., total test score) through the judgmental analysis in practice. Also, if a test has no biased items, then we should expect perfect unreliability in the detection of bias (i.e., all significant results would be Type I error).

One solution to this problem is that homogeneous or valid subtests from the total test should be selected, and that the subtest score should be used as the matching variable (Hambleton et al., 1993). Hambleton et al argued that this provided a purified and more valid criterion for matching the Reference and Focal groups. These subtests may be identified either statistically, or by judgmental analysis (Hambleton et al., 1993). However, because of the characteristics of performance assessment (i.e., limited number of items), this may not be applicable in practice.

In conclusion, a careful judgmental review that comes before DIF analysis might improve the degree of purification of the matching criterion. Specifically, inflated Type I error that resulted from the contamination of the matching variable can be reduced or minimized through the "sensitivity reviews (Ramsey, 1993)".

Limitations and Recommendations for Future Research

This study utilized model-generated data rather than real data. The baseline performance data was established for the procedures without being confounded by factors which are likely to vary in practice. Thus, several assumptions were made to simulate ideal testing condition under which the methodology of the procedures could be examined.

The first assumption was made when the Integer Transformed Theta (ITT) scores were used as the matching variable. The comparability of the Reference and Focal groups was achieved by matching them on the basis of a measure of test performance. In the second component of this simulation study, the ITT created a matching criterion that was free from DIF. However, this condition would be very difficult to obtain using real data. Thus, there is a need to develop a criterion variable which is purely unbiased.

In order to use an internal matching criterion, a purified subtest score proposed by Hambleton and Others (1993) may be used as a criterion for

matching the Reference and Focal groups. Regarding an external matching criterion, the standardization approach (See, Schmitt, Holland, & Dorans, 1993) may be applicable in performance assessments.

The second assumption was brought about when the condition of the average magnitude of DIF items in a given test was the same as in the second test, the overall effect size of the DIF items on the total test score was directly proportional to the number of DIF items in a test. However, it was not determined whether any differences existed between the effect of two DIF items of 0.15 and 0.25 magnitudes and the effect of two items of 0.20 and 0.20 magnitudes. Further research should examine the possible differences between these two conditions.

In particular, when unequal ability distributions between two groups existed, the relatively small influence of the number of DIF items in the matching variable for identifying nonuniform DIF was found. If this condition is close to practice, further research in this area is needed along the lines of this study.

The third assumption concerned the design of nonuniform DIF. A nonuniform DIF condition, under which positive and negative DIF cancel each other entirely, was designed to examine the relative statistical power of the three statistics. However, this ideal nonuniform DIF condition might be rare in practice. Thus, it is strongly recommended to study the effect of the degree of

nonuniform DIF (i.e., the magnitude of how much positive and negative DIF cancel each other) for detecting DIF in polytomously scored items.

As discussed earlier, the effect of the number of DIF items on matching criterion appeared to be stronger for detecting uniform DIF than identifying nonuniform DIF. Then, the following questions are raised. Which type of DIF exists most in practice? Uniform DIF dominant, nonuniform DIF dominant, or both of these? What extent of the proportion of DIF items on the matching criterion is tolerable for detecting DIF related to the type of DIF? Future studies along this line are recommended.

Lastly, why do these three methods perform differently? Both LDFA and MH are based on a nonparametric test, specifically the loglinear model, while CT is based on a parametric test. The results of the generalized MH estimation for the detection of uniform DIF in the polytomously scored items were essentially the same as those of the LDFA. The model of the MH is similar to the model of the LDFA when the item score-by-ability interaction is not included in the model.

However, the results show that the empirical powers of the two methods are not identical for uniform DIF. This difference in DIF measurement between the two methods seems to be related to the metric in which the statistic is portrayed. The delta metric has been the metric of choice for the MH method, while the metric used by the LDFA has been the p-metric (Dorans & Holland,

1993). In fact, the two methods are measuring essentially the same thing, DIF, in slightly different ways. Therefore, further research into a mathematical explanation of why these three methods exhibit different statistical powers is recommended.

Summary

Obviously, the procedure of the logistic discriminant function analysis appeared the most promising and powerful for detecting DIF on polytomously scored items. However, some of the practical limitations of the statistic are clear. First, the selection of a matching variable is a significant decision. The most closely matched reliable criterion should be employed for matching the groups.

Second, whenever possible, relatively large sample sizes should be used. However, when very large samples are used, it may be important to use measures of both statistical significance and magnitude in examining items. Because the power of the statistic increases with sample size, trivial levels of DIF may be identified as statistically significant. Hambleton and Others (1993) suggested using a system such as that currently in use at Educational Testing Service (see Chapter 2) when samples are in excess of 1,000 per group. Since the study for effect size is not included in this study, further research into the magnitude of DIF for the three methods is recommended.

It was found in this study that unequal mean ability between groups is helpful to increase the empirical powers of the LDFA, MH, and CT for nonuniform DIF. However, the relationship between the degree of unequal mean ability and the degree of the statistical powers of the procedures was not revealed.

Also, it should be remembered that DIF analysis is only one component of the extensive research for the validity and fairness of performance assessment, while it is essential to the appropriateness of test use for subgroups affected by testing. Camillie and Shepard (1994) maintained that DIF analysis cannot be applied mechanically "because significant DIF is not synonymous with bias (p. 155)."

Therefore, DIF items need to be carefully reviewed and studied in relation to their relevance to the intended test construct through judgmental analysis. It is necessary to use both the DIF statistical method and the judgmental method in order to create the first step for testing situations which are truly free from bias.

Finally, it should be emphasized that the studies of DIF should not be treated as if they address broader questions of test bias and fairness. As mentioned above, DIF analysis is only one component of the extensive research that is necessary for the validity and fairness of a test. Therefore, we should not maintain that a test is unbiased and fair only because it has been examined through DIF analyses.

BIBLIOGRAPHY

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley.

Airasian R. W. (1991). Classroom Assessment. New York: McGraw-Hill, Inc..

Allen, N. L., & Holland, P. W. (1993). A model for missing information about the group membership of examinees in DIF studies. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 241-254). Hillsdale, NJ: Lawrence Erlbaum.

American Educational Research Association, and American Psychological Association, National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, D.C.: American Psychological Association.

Andersen, E. B. (1983). A general latent structure model for contingency table data. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement. (pp. 117-138). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 96-116). Baltimore, Maryland: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18(1), 59-62.

Begg, C. B. and Gray, R. (1984). Calculation of polytomous logistic regression parameters using individualized regressions. Biometrika, 71(1), 11-18.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille edition. Journal of Educational Measurement, 26(1), 67-79.

Berlak, H., Newmann, F. M., Adams E., Archbald , D. A., Burgess, T., Raven, J., & Romberg, T. A. (1992). Toward a new science of educational testing and assessment. Albany, New York: State University of New York Press.

Bock, R. D. (1993). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 115-122). Hillsdale, NJ: Lawrence Erlbaum.

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 277-279). Hillsdale, NJ: Lawrence Erlbaum.

Bull, S. B. and Donner, A. (1987). The efficiency of multinominal logistic regression compared with multiple group discriminant analysis. Journal of the American Statistical Association, 82, 1118-1122.

Burrill, L. E. (1982). Comparative studies of item bias methods. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 161-179). Baltimore, Maryland: Johns Hopkins University Press.

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 321-335). Hillsdale, NJ: Lawrence Erlbaum.

Calfee, R. C., & Perfumo, P. (1993). Student portfolios: Opportunities for a revolution in assessment. Journal of Reading, 36, 532-537.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. Applied Psychological Measurement, 16(2), 129-147.

Camilli, G. (1993). The case against DIF techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 397-417). Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G., & Shepard, L. A. (1994). Methods for Identifying Biased Test Items. Newbury Park, CA: Sage Publications.

Clauser, B., Mazor, K., & Hambleton, R. K. (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 15(4), 353-359.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using Mantel-Haenszel procedure. Applied Measurement in Education, 6(4), 269-279.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. Journal of Educational Measurement, 31(1), 67-78.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cleary, T. A. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Cohen, A. S., Kim, S., & Subkoviak, M. J. (1991). Influences of prior distributions on detection of DIF. Journal of Educational Measurement, 28(1), 49-59.

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's χ^2 and Raju's Area Measures in Detection of DIF. Applied Psychological Measurement, 17(1), 39-52.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Eds.), Educational Measurement (Part 1, pp. 201-219). New York: Macmillian Publishing Company.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 25-29). Hillsdale, NJ: Lawrence Erlbaum.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Forth Worth, TX: Harcourt Brace Jovanovich College Publishers.

Darlington, R. B. (1971). Another look at "cultural fairness." Journal of Educational Measurement, 8, 71-82.

Demaris, A. (1992). Logit Modeling. Newbury Park, CA: Sage Publications.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus Thick matching in the Mantel-Haenszel procedure for detecting DIF. Journal of Educational Statistics, 18(2), 131-154.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24(2), 157-166.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2, 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. Journal of Educational Measurement, 23(4), 355-368.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement, 29(4), 309-319.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Ellis, B. B.(1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74(6), 912-921.

Englehard, G., Anderson, D., & Gabrielson, S. (1990). An empirical comparison of Mantel-Haenszel and Rasch procedures for studying differential item functioning on teacher certification tests. Journal of Research and Development in Education, 23, 172-179.

Finch, F. L., & Dost, M. A. (1992, June). Toward an Operational Definition of Educational Performance Assessments. Paper presented at the Assessment Conference of the Education Commission of the States/Colorado Department of Educational Assessment, Boulder, CO. (ERIC Document Reproduction Service No. ED 353 287)

Fowler, R. L., & Clingman, J. M. (1992). Identifying negatively discriminating items when test scores are not normally distributed. Educational and Psychological Measurement, 52(1), 31-39.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of differential item functioning for black and white examinees. Journal of Educational Measurement, 27(4), 329-343.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. Journal of Educational Measurement, 26(2), 147-160.

Hambleton, R. K. (1994, July). Methods of setting standards on performance assessments in state-wide assessment contexts: a proposal. Paper presented at the meeting of the meeting of State Collaborative on Assessment and Student Standards, Traverse City, MI.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. European journal of Psychological Assessment, 9(1), 1-18.

Hambleton, R. K. & Jones, R. W. (1994). Comparison of Empirical and Judgmental Methods for Detecting Differential Item Functioning. Educational Research Quarterly, 18(1), 21-36.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory, Newbury Park, CA: Sage Publications.

Harman, D. (1980). On traditional testing. In E. L. Baker & E. S. Quellmalz (Eds.), Educational testing and evaluation (pp.229-236). Beverly Hills, CA: Sage Publications.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunter, J. E. (1975, December). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore, Maryland: Johns Hopkins University Press.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? Harvard Educational Review, 39, 1-123.

Jensen, A. R. (1980). Bias in Mental Testing. New York: Free Press.

Judd, C. M., & McClelland, G. H. (1989). Data analysis: model comparison approach. San Diego, CA: Harcourt Brace Jovanovich.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27(4), 307-327.

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. Journal of Educational Measurement, 29(1), 51-66.

Kim, S., & Cohen, A. S. (1991). A comparison of two measures for detecting differential item functioning. Applied Psychological Measurement, 15(3), 269-278.

Klieme, E., & Stumpf, H. (1991). DIF: A computer program for the analysis of differential item performance. Educational and Psychological Measurement, 51(3), 669-671.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 317-319). Hillsdale, NJ: Lawrence Erlbaum.

Linn, R. L. (1976). In search of fair selection procedures. Journal of Educational Measurement, 13(1), 53-58.

Linn, R. L. (1993). The use of differential item functioning statistics: a discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 349-366). Hillsdale, NJ: Lawrence Erlbaum.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18(2), 109-118.

Longford, N. T., Holland, P. W., and Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 171-196). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Masters, G. N. (1988). Partial credit models. In J. Keeves (Ed.), Educational research, methodology, and measurement, an international handbook. New York: Pergamon.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement, 54(2), 284-291.

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 389-396). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), Educational Measurement (Part 1, pp. 13-103). New York: Macmillian Publishing Company.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. Journal of Educational Measurement, 30(2), 107-122.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. Applied Psychological Measurement, 16(4), 381-388.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. Applied Psychological Measurement, 16(4), 389-402.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. Journal of Educational Measurement, 30(4), 293-311.

North Central Regional Educational Laboratory. (1994). State student assessment program database 1993-94. Oak Brook, IL: North Central Regional Educational Laboratory.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16(3), 237-248.

Oshima, T. C., Raju, N. S., & Flowers, C. (1993, April). Comparison of Empirical and Judgmental Methods for Detecting Differential Item Functioning. Paper presented at the annual meeting of American Educational Research Association, Atlanta. (ERIC Document Reproduction Service No. ED 365 707)

Osterlind, S. J. (1989). Constructing test items. Norwell, MA: Kluwer Academic Publishers.

Osterlind, S. J. (1983). Test item bias. Beverly Hills, CA: SAGE Publications, Inc..

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13(1), 3-39.

Plake, B. S., Patience, W. M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: effect of item arrangement. Educational and Psychological Measurement, 48, 885-894.

Popham, W. J. (1990). Modern educational measurement: A practitioner's perspective (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Pulliam, J. D. (1991). History of education in America. New York: Macmillian Publishing Company.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2(1), 1-13.

Ramsay, J.O. (1993). Comments on the Monte Carlo study of Donoghue, Holland, and Thayer. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 167-169). Hillsdale, NJ: Lawrence Erlbaum.

Ramsey, P. A. (1993). Sensitivity review: the ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 199-227). Baltimore, Maryland: Johns Hopkins University Press.

Rogers, H. J. and Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential Item functioning. Applied Psychological Measurement, 17(2), 105-116.

Rothman, R. (1995). Measuring Up. San Francisco: Jossey-Bass Publishers.

Ryan, K. E. (1991). The performance of the mantel-Haenszel Procedure across samples and matching criteria. Journal of Educational Measurement, 28(4), 325-337.

Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. Psychometrika, 58(1), 119-138.

Samejima, F. (1993). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. Psychometrika, 58(2), 195-209.

Sawyer, R. L., Cole, N. S., & Cole, J. W. L. (1976). Utilities and the issue of fairness in a selection theoretic model for selection. Journal of Educational Measurement, 13(1), 59-76.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16(3), 143-152.

Scheuneman, J. D. (1981). A response to Baker's criticism. Journal of Educational Measurement, 18(1), 63-66.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 180-198). Baltimore, Maryland: Johns Hopkins University Press.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24(2), 97-118.

Schmeiser, C. B. (1982). Use of experimental design in statistical item bias studies. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 64-95). Baltimore, Maryland: Johns Hopkins University Press.

Schmitt, A. P. and Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27(1), 67-81.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58(2), 159-194.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Shepard, L. A. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 9-30). Baltimore, Maryland: Johns Hopkins University Press.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22(2), 77-106.

Steinberg, L., Thissen, D., & Wainer, H. (1990). Validity. In H. Wainer (Eds.), Computerized Adaptive Testing: A Primer (pp. 187-230). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Stiggins, R. J. (1994). Student-centered classroom assessment. New York: Macmillan.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. Applied Measurement in Education, 4, 263-273.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22(4), 271-286.

Swaminathan, H. and Rogers, H. H. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Tang, Huixing (1994, April). A Simultaneous Approach to Multi-Factor DIF Analysis, Paper presented at the 1994 annual meeting of the National Council on Measurement in Education, New Orleans.

Tang, Huixing (1994, April). Step Fit Analysis with Polytomously Scored Items, Paper presented at the 1994 annual meeting of the American Educational Research Association, New Orleans.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. Journal of Educational Measurement, 25(4), 301-319.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test Validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item, Journal of Educational Measurement, 26(2), 161-176.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Thorndike, R. L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8, 63-70.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Eds.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore, Maryland: Johns Hopkins University Press.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: definitions and detection. Journal of Educational Measurement, 28(3), 197-219.

Waller, M. I. (1981). A procedure for comparing logistic latent trait models. Journal of Educational Measurement, 18(2), 119-125.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. Applied Measurement in Education, 6(1), 1-19.

Westers, P. (1993). The solution-error response-error model: a method for the examination of test item bias. Unpublished doctoral dissertation, Twente University, The Netherlands. (ERIC Document Reproduction Service No. ED 366 642)

Wiersma, W., & Jurs, S. G. (1990). Educational measurement and testing (2nd ed.). Needham Heights, MA: Allyn and Bacon.

Wigdor, A. K., & Garner, W. R. (1982). Ability testing: uses, consequences, and controversies. Washington, DC: National Academy Press.

Wiggins, G. P. (1993). Assessing student performance: Exploring the purpose and limits of testing. San Francisco: Jossey-Bass Publishers.

Wiggins, G. P. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

William, M. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525-543.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30(3), 233-251.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26(1), 55-66.

VITA

The author of this dissertation, Hae-Seong Park, graduated with a Doctor of Philosophy degree from Louisiana State University in December of 1995. He concentrated his graduate study in educational research and statistics. While serving as a graduate assistant at LSU he taught the graduate statistics lab and assisted in research projects.

In addition to his Ph.D., the author has earned a Bachelor of Arts and a Master of Divinity from Chongshin College; and a Master of Christian Education and a Master of Theology from the Reformed Theological Seminary. He is a certified secondary school teacher in Korea and he taught Ethics in the Hwanil Senior High School for two years. He was ordained as a minister in 1983 and served as a chaplain in Korean Army for five years. He has helped the Starkville Korean Church, the Hattiesburg Mission Church, and the Lafayette Korean Church as a part time pastor from 1989 to the present.

The author is presently employed as a psychometrician at the Office of Research and Development in the Louisiana Department of Education. His work is concentrated on analysis of statewide test data.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Hae-Seong Park

Major Field: Educational Research

Title of Dissertation: Differential Item Functioning in Performance Assessments: A Comparison of Three Procedures

Approved:

Eugene Kennedy
Major Professor and Chairman

John W. Lasher
Dean of the Graduate School

EXAMINING COMMITTEE:

Charles Tuddle

Abdas Tavakkol

S. Kim MacGregor

Patsy W. Perritt

Date of Examination:

September 27, 1995